




3 1761 10374369 6



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743696>

12-001



Government
Publications

139

SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005

•

VOLUME 31

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005 • VOLUME 31 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2005

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

July 2005

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Members J. Gambino
J. Kovar
H. Mantel

Past Chairmen G.J. Brackstone
R. Platek

E. Rancourt (Production Manager)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Deputy Editor H. Mantel, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*

D.A. Binder, *Statistics Canada*

J.M. Brick, *Westat, Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J.L. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

M.A. Hidirolou, *Office for National Statistics*

G. Kalton, *Westat, Inc.*

P. Kott, *National Agricultural Statistics Service*

J. Kovar, *Statistics Canada*

P. Lahiri, *JPSM, University of Maryland*

G. Nathan, *Hebrew University*

D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

M. Thompson, *University of Waterloo*

Y. Tillé, *Université de Neuchâtel*

R. Valliant, *JPSM, University of Michigan*

V.J. Verma, *Università degli Studi di Siena*

J. Waksberg, *Westat, Inc.*

K.M. Wolter, *Iowa State University*

A. Zaslavsky, *Harvard University*

Assistant Editors

J.-F. Beaumont, P. Dick and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, Dr. M.P. Singh, singhmp@statcan.ca (Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.



Survey Methodology

A journal Published by Statistics Canada

Volume 31, Number 1, June 2005

Contents

In This Issue	1
M. Winglee, R. Valliant and F. Scheuren A Case Study in Record Linkage	3
D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski and R. Mallick The Effect of Record Linkage Errors on Risk Estimates in Cohort Mortality Studies	13
Jan A. van den Brakel and Robbert H. Renssen Analysis of Experiments Embedded in Complex Sampling Designs	23
Takahiro Tsuchiya Domain Estimators for the Item Count Technique	41
Marco Di Zio, Ugo Guarnera and Orietta Luzi Editing Systematic Unity Measure Errors Through Mixture Modelling	53
Wai Fung Chiu, Recai M. Yucel, Elaine Zanutto and Alan M. Zaslavsky Using Matched Substitutes to Improve Imputations for Geographically Linked Databases	65
Balgobin Nandram and Jai Won Choi Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES Data	73
Mingue Park and Wayne A. Fuller Towards Nonnegative Regression Weights for Survey Samples	85
Short Notes	
Per Gösta Andersson and Daniel Thorburn An Optimal Calibration Distance Leading to the Optimal Regression Estimator	95
Peter Lynn and Siegfried Gabler Approximations to b^* in the Prediction of Design Effects Due to Clustering	101
Jane L. Meza and P. Lahiri A Note on the C_p Statistic Under the Nested Error Regression Model	105

In This Issue

This issue of *Survey Methodology* is dedicated to Gordon J. Brackstone, who recently retired from Statistics Canada. He was Assistant Chief Statistician for the Informatics and Methodology field and had been chairman of the *Survey Methodology* management board since 1987. His continuous support to the journal has been marked by great insight and motivated by a constant desire to foster high standards of methodology practices. Further, he also authored several articles that appeared in the journal. We wish to express our extreme gratitude to Gordon J. Brackstone.

The current issue contains eight regular papers on a variety of topics, and three short communications. As mentioned in the previous issue of the journal, we are introducing a new Short Communications section in *Survey Methodology*. This section will contain shorter papers, typically around four pages. Possible topics of short communications include presentation of new ideas without the full development of a regular paper, brief reports of empirical work, and discussions or supplements related to other papers published in the journal.

For the past four years the June issue of *Survey Methodology* has included an invited paper in honour of Joseph Waksberg. Starting this year, this annual invited paper will be published in the December issue of the journal, bringing it more in line with the associated Waksberg address delivered at Statistics Canada's annual methodology symposium in the autumn. The author of this year's Waksberg paper is J.N.K. Rao and his paper will be on the "Interplay Between Sample Survey Theory and Methods: an Appraisal".

In the opening paper of this issue, Winglee, Valliant and Scheuren present a new simulation approach to estimation of error rates for threshold selection in record linkage. For each potential matched pair there is a vector of comparison outcomes that determines the linkage weight. A multinomial model is assumed for each comparison outcome, with different multinomial distributions for true matches and true non-matches. The distributions are estimated from a sample, and then used to simulate the distributions of the linkage weights for true matches and true non-matches. The method is illustrated in a case study using data from the U.S. Medical Expenditure Panel Survey (MEPS).

Krewski, Dewanji, Wang, Bartlett, Zielinski and Mallick investigate the effects of record linkage errors, both false positives and false negatives, on risk estimates in cohort studies. They show analytically how linkage errors introduce both bias and additional variability into observed and expected numbers of deaths, as well as into estimates of standardized mortality ratios and relative risk regression coefficients. They discuss their results in their conclusions, and point to further work that needs to be done in this area.

The paper by van den Brakel and Renssen addresses the problem of testing hypotheses between different survey implementations, such as different questionnaire designs, when a complex sampling design is used. A design-based theory is developed for cases where the survey implementations are assigned to subsamples through completely randomized experimental designs or randomized block experimental designs. The theory also makes use of measurement error models. Design-based Wald statistics are used to compare the different survey implementations.

Tsuchiya approaches the long-standing problem of asking respondents sensitive questions in an interesting fashion. Instead of using the randomized response approach that allows little control for the researcher, he proposes that the item count technique be adapted for sensitive questions. The item count technique presents the respondent with a list of several phrases, from which the respondent selects all that apply to him. The researcher constructs the list in two ways: the first list contains the sensitive phrase while the second list does not. Tsuchiya presents various estimators for this technique and gives an interesting example related to the Japanese national character.

In the paper by DiZio, Guarnera and Luzi, finite mixture models are used to detect errors that are due to an incorrect unit of measurement at the collection stage of the survey. In a multivariate context and assuming that the data are multivariate normal, the procedure can identify which variables are in error for a given sampled unit. The authors also provide diagnostics for prioritizing cases to be investigated more deeply through clerical review. The proposed methodology is illustrated through an example with simulated data and an example with real data.

Chiu, Yucel, Zanutto and Zaslavsky present a method for multiple imputation of missing contextual variables for use in regression analysis. For each record missing the variable, and for a sample of complete records, matched cases are selected based on a set of matching variables. The sample of complete records is then used to estimate a regression adjustment for other variables not included among the matching variables. The contextual variables for the incomplete records are then multiply imputed. The authors then show an application to a colorectal cancer study, and use simulations to compare their approach to three other nonresponse adjustment methods.

Nandram and Choi examine the important problem of nonignorable nonresponse in small-area estimation of a health status variable. When confronted with an example where the usual estimators are biased because of the excessive number of nonrespondents, they attempt to account for the differences through modeling. Nandram and Choi use two nonignorable nonresponse hierarchical Bayes models, a selection model and a pattern model, to analyze the health data. An important consideration to their modeling is the incorporation of the input from doctors concerning the nonresponse pattern and the outcome variable. The results give an accurate non-response adjustment and a better measure of precision.

Park and Fuller propose a method to reduce the probability of obtaining negative estimation weights when using a regression estimator. Their method consists of first approximating inclusion probabilities, conditional on Horvitz-Thompson estimates for a vector of auxiliary variables, and then using these approximate conditional inclusion probabilities as initial weights in a regression estimator. Their method is shown to work well in a simulation study. The weights obtained from this method are also compared to weights from quadratic programming, the raking ratio, the logit procedure and maximum likelihood.

In the first of three short communications included in this issue, Andersson and Thorburn show that the optimal regression estimator can be expressed as a calibration estimator with an appropriately chosen distance function. The resulting optimal estimator is asymptotically more efficient than the usual Generalized Regression (GREG) estimator. A small simulation study illustrates several situations where the optimal estimator is significantly more efficient than the GREG estimator.

Lynn and Gabler extend the results of Gabler, Hader and Lahiri (volume 25, 1999) on Kish's expression for the design effect due to clustering. They give a practical approach to estimating Kish's quantity at the sample design stage when only the total numbers of observations and of clusters are needed.

Meza and Lahiri examine the limitations of a standard regression model selection criterion, Mallows' statistic, for nested error regression models. They show, that while a straightforward application of Mallows' statistic may result in inefficient model selection methods, a suitable transformation of the data may be the answer.

Finally, we would like to inform you that Harold Mantel will now hold the new position of Deputy Editor. Harold has been part of the Editorial Board for the last 15 years. His dedication to the journal has been notable and his continuous involvement in the editorial process has been instrumental in ensuring that *Survey Methodology* remains a high quality publication.

M.P. Singh

A Case Study in Record Linkage

M. Winglee, R. Valliant and F. Scheuren¹

Abstract

Record linkage is a process of pairing records from two files and trying to select the pairs that belong to the same entity. The basic framework uses a match weight to measure the likelihood of a correct match and a decision rule to assign record pairs as "true" or "false" match pairs. Weight thresholds for selecting a record pair as matched or unmatched depend on the desired control over linkage errors. Current methods to determine the selection thresholds and estimate linkage errors can provide divergent results, depending on the type of linkage error and the approach to linkage. This paper presents a case study that uses existing linkage methods to link record pairs but a new simulation approach (SimRate) to help determine selection thresholds and estimate linkage errors. SimRate uses the observed distribution of data in matched and unmatched pairs to generate a large simulated set of record pairs, assigns a match weight to each pair based on specified match rules, and uses the weight curves of the simulated pairs for error estimation.

Key Words: File matching; Linkage error rates; Match weight; Selection threshold; Medical records.

1. Introduction

The basic record linkage framework by Newcombe Kennedy, Axford and James (1959) and Fellegi and Sunter (1969) uses a match weight to measure the likelihood of a correct match and a decision rule to classify record pairs. The optimal decision rule uses two match weight thresholds for selection (an upper threshold above which a link is treated as a match and a lower threshold below which a link is treated as a nonmatch). The choice of these thresholds depends on the acceptable pre-set linkage error rate and the requirement to minimize the number of links with indeterminate status between the two thresholds. Nowadays, practitioners of computerized linkage systems often use a single selection threshold to avoid manual intervention of the indeterminate links. Linkage decisions are typically made automatically after the system is "tuned" to achieve pre-set error levels. The challenge is that current methods to determine the selection threshold and to estimate linkage errors can produce divergent results depending on the type of linkage error, the choice of comparison space, and the estimation method.

This paper shares our experience with fellow practitioners who need a method to guide linkage selection and error estimation. Our case study used medical event files from the US Medical Expenditure Panel Survey (MEPS). MEPS collects medical expenditure data from both household respondents and their medical providers. The purpose is to combine the data from both sources for supporting annual estimations of medical utilization and expenditures (see Agency for Healthcare Research and Quality 2001 for more details on MEPS).

Here we discuss the linkage with three sets of annual medical event files – MEPS 1996, MEPS 1997, and MEPS 1998. Each set consisted of a household file containing events reported by household respondents for a given year and a medical provider file containing the corresponding events reported by medical providers of the household respondents. On average, approximately 50,000 medical events were reported for close to 10,000 persons, and around 15,000 person-provider units each year.

We used two model-based alternatives for linkage error estimation. One of these uses simulation to develop a distribution of the weights for various levels of agreement. This technique, called SimRate, begins by generating weight distributions for matched and unmatched record pairs. Using these, SimRate can then provide estimates of linkage error rates for different threshold levels. The error rates can then be used as a guide to action and a way to measure success. SimRate is contrasted with a second modeling approach created by Belin and Rubin (1995). As we hope to show, there is a role for both approaches; each has strengths as illustrated in the comparisons.

2. Mixture Models and Simrate Approaches

The mixture modeling method of linkage error estimation, as presented in Belin and Rubin (1995), has several attractive features. It is flexible in a sense that the weight creation process does not have to be considered directly. Hence, this method can be applicable to many different ways of creating weights. Once a model is specified, error

1. M. Winglee, Westat, Statistical Group, 1650 Research Boulevard, Rockville, MD 20850-3195, U.S.A.; R. Valliant, Joint Program for Survey Methodology, University of Maryland and University of Michigan; F. Scheuren, NORC, University of Chicago.

rates can be examined for a continuum of potential threshold values and confidence bands can be constructed to monitor the precision of error estimates (see section 7).

Mixture modeling does have limitations. While the method provides a particular kind of error rate – the proportion of linked records that are actually unmatched pairs, overall false positive and false negative error rates cannot be estimated since nonlinked pairs are not considered. The error rate that is estimated is conditional on the set of linked pairs of records. Furthermore model parameters may be hard to estimate if the weight distributions for the matched and unmatched sets are not separable (see Winkler 1994).

A key assumption in the Belin–Rubin approach is that it is possible to transform the distributions of the weights in the matched and unmatched sets to make them normal. Now a real difficulty exists here in that the transformed weights may be far from normal when the weight distribution for either the matched or unmatched sets is multimodal.

Another critical requirement is to have a training data set whose characteristics are very similar to those that are to be matched. Without a good training data set, the input parameter estimates for the mixture model may be poor, affecting the final estimated error rates obtained. Based on our application using annual medical event data repeated over three years, the parameters were not stable over time. This instability necessitated a training set for each year, making the Belin–Rubin approach impractical in our application because of the cost and time it required.

The simulation approach, SimRate, like mixture modeling, has the ability to examine different thresholds, allowing the user to monitor both the sensitivity and specificity of the decision rule for selecting linked pairs. As long as the process used to create match weights can be realistically modeled, customized methods of weight assignment like the one used in the current case study can be accommodated. The method does require the generation of pairs of records using the distribution of characteristics for the matched and unmatched sets. Some effort is needed to realistically generate the populations of pairs. In our work we have been successful with multinomial models for generating these populations.

3. Threshold Weight and Linkage Error Estimation

Several methods are available in the literature for selecting true matches and for estimating linkage errors (e.g., Bartlett, Krewski, Wang and Zielinski 1993, Armstrong and Mayda 1993, Belin 1993, Belin and Rubin 1995 and Winkler 1992, 1995). See Fellegi (1997) for an overview of evolutions in record linkage, Tepping (1968) and Larsen and Rubin (2001) for other linking methods, and

Scheuren (1983) for a capture-recapture method to estimate omission error.

Comparison of estimates from the different approaches is complicated by the fact that each approach tends to focus on different error components. In fact, the methods used in the linkage literature to construct linkage error rates are somewhat inconsistent. We illustrate this problem below.

Table 1 shows a 2×2 contingency table tabulating the numbers of true matched and unmatched pairs and declared linked and nonlinked pairs selected by linkage systems. Estimates of linkage error rates can be constructed relative to the true totals shown in the columns. An estimate of false positive linkage error rate under the Fellegi and Sunter framework is $\hat{\mu} = P(A_1 | U) = n_{12} / n_{\bullet 2}$ and that of false negative linkage error rate is $\hat{\lambda} = P(A_3 | M) = n_{21} / n_{\bullet 1}$ (see also Armstrong and Mayda 1993). These are the rates that SimRate is designed to estimate. They answer the question – “Of the set of true matched (or unmatched) pairs, what proportion is not correctly identified?”

Table 1
A Contingency Table for Evaluating Linkage Errors

Declared set	True set		Declared total
	Match (M)	Unmatch (U)	
Link (A_1)	n_{11} true positive	n_{12} false positive	$n_{1\bullet}$
Nonlink (A_3)	n_{21} false negative	n_{22} true negative	$n_{2\bullet}$
True total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Some linkage evaluations have also considered rates relative to the declared totals in the rows. For instance, Gomatam, Carter, Ariet and Mitchell (2002) used $n_{12} / n_{1\bullet}$ and labeled it the positive predictive power of the linkage system. Others, however, have labeled this as the false match rate (Belin and Rubin 1995) or false positive declared rate (Bartlett *et al.* 1993). Rates constructed in this manner answer the question – “Of the declared linked (or nonlinked) pairs, what proportions are wrong?” Both questions are important in selecting matched pairs and should be addressed. That is one of the appeals in employing both SimRate and Belin–Rubin, if possible.

4. Simrate Weight Distribution Methods to Estimate Linkage Error

How to best estimate the linkage errors, given a limited budget and time schedule, is a difficult question. Accurate estimation of linkage errors should depend on at least two factors – the power of the identifying fields to unambiguously identify events that are true matches and the linkage method used. Taken together it is then possible, in a given setting, to specify linkage categories, estimate agreement probabilities, and determine match weights.

Following Newcombe and Kennedy (1962) and Jaro (1989), we adopt a weight distribution approach in our application that can take all these factors into consideration. The basic step is to first compute the match weight and order all possible configurations of agreement and disagreement outcomes of the comparison fields by match weight. Then we plot the cumulative distribution function of the weights for matched and unmatched pairs, and use the resulting weight chart to determine thresholds to attain desired levels of false positive and false negative error rates.

An ideal method to develop these curves might be to begin with a set of record pairs for which the truth is known. If resources are available, we could use a large set of true matched pairs, order them by match weight, and observe what proportion is above or below a given threshold. Similarly, we could take a large set of pairs, known to be true unmatched pairs, order them by weights, and again tabulate the proportion on either side of the threshold. The proportion of true matched pairs with weights below the threshold and the proportion of true unmatched pairs with weights above the threshold would then be estimates of the error rates associated with the way in which the matching algorithm is implemented.

One method to approximate this "ideal" approach (see also Bartlett *et al.* 1993) is to sample record pairs and use manual review to determine the true match status. Once the true pairs are known, we can attach the match weights from whatever linkage system is being used and then develop cumulative weight distributions, as discussed above. This method is, of course, subject to the well-known time and other resource limitations of manual review and is seldom practical with a large sample.

An alternative method is to generate the cumulative weight distributions through simulation. That is the heart of the SimRate approach. To explain in some detail, denote a record pair by r and a comparison field by v ($v = 1, \dots, V$ fields). The comparison outcome situations in our application included partial agreements and multiple outcome categories beyond the basic agreement and disagreement categories (see also Newcombe 1988). Therefore, we denote that each field v has $i = 1, \dots, c_v$ outcome categories. The outcome indicator is $\mathbf{y}_{rv} = (y_{rv1}, \dots, y_{rv c_v})$, a vector of indicators showing the category into which pair r falls. One of the values of $y_{rv i}$ will be 1 and the others 0 for each field.

The particular theory supporting the SimRate approach is to assume that \mathbf{y}_{rv} has one multinomial distribution if pair r is a matched pair and a different multinomial distribution if it is an unmatched pair. We can then model the \mathbf{y}_{rv} vectors as having a multinomial distribution with parameters $\mathbf{m}_v = (m_{v1}, \dots, m_{v c_v})$ if the pair is a matched pair and parameters $\mathbf{u}_v = (u_{v1}, \dots, u_{v c_v})$ if the pair is an

unmatched pair. Then the probability $m_{vi} = P(\text{field } v \text{ category } i \text{ agrees in pair } r | r \in M)$ is the conditional probability of agreement for field v category i , given that the record pair r is in the set M of true matched pairs. In contrast, the probability $u_{vi} = P(\text{field } v \text{ category } i \text{ agrees in pair } r | r \in U)$ is the conditional probability of agreement for field v category i , given that the record pair r is in the set U of true unmatched pairs. Assuming independence of the matching variables, $v = 1, \dots, V$, we can specify the joint probability of $\mathbf{y}_r = (y_{r1}, \dots, y_{rV})$ if a pair r is a match, as

$$P(\mathbf{y}_r | r \in M) = \prod_{v=1}^V \prod_{i=1}^{c_v} m_{vi}^{y_{rv i}}.$$

The corresponding probability of the same configuration of data, if the pair is really an unmatched pair, is

$$P(\mathbf{y}_r | r \in U) = \prod_{v=1}^V \prod_{i=1}^{c_v} u_{vi}^{y_{rv i}}.$$

SimRate uses Monte Carlo simulation methods to generate a large number of realizations of matched pairs and unmatched pairs using estimates of the probabilities m_{vi} and u_{vi} . For each simulated pair, a match weight w_r , which applies to a given configuration of data, is calculated. For a given realization \mathbf{y}_r , a weight w_r is computed for the pair by summing the weights for the randomly generated categories that the pair fell into. The match weight w_r of a record pair is typically estimated as

$$w_r = \log_2 \frac{\prod_{v=1}^V \prod_{i=1}^{c_v} m_{vi}^{y_{rv i}}}{\prod_{v=1}^V \prod_{i=1}^{c_v} u_{vi}^{y_{rv i}}}.$$

See section 6 on the match weights used in our simulation.

The cumulative distribution of these weights for the simulated matched pairs is then plotted as "Sim- M ". Similarly, the reverse cumulative distribution for the unmatched pairs is plotted to generate "Sim- U " (see Figure 1, section 8, for an example of the simulation curves used in this study). The simulated proportion of matched pairs whose weights are below the cutoff is the estimate of the false negative error rate. The simulation proportion of unmatched pairs whose weights are above the cutoff is the estimate of the false positive error rate.

This approach requires that empirical estimates be made of the distributions among the matching variables of both true matched and true unmatched pairs. Even though the weight algorithm may involve the assumption of independence among matching variables, the actual data may show dependence. As long as artificial pairs can be generated that realistically follow the observed distribution of the data (incorporating any dependencies), then this method should provide suitable error rate estimates.

In our case study, we modeled data fields as having independent multinomial distributions, but this may not be reasonable in other applications. The SimRate concept can apply to any algorithm where weights and a cutoff point are used for classification. Thus, methods other than Fellegi and Sunter (1969), like Belin and Rubin (1995), might also be evaluated in this way. If methods are needed to deal with dependent categorical variables, the multivariate multinomial distributions in Johnson, Kotz, and Balakrishnan (1997, Chapter 26) may be appropriate. However, in applications similar to ours, the simplest procedure for accounting for dependence is to form cross-classifications of the variables that are related and to estimate probabilities for each cell in a cross-table. For example, if two variables with c_1 and c_2 categories are associated, then we can estimate the joint probability, p_{ij} , for each cell in the $c_1 \times c_2$ table and use those in the simulation. Sparse data will naturally limit the number of cells for which this is feasible. But in the presence of sparse data, the penalty for model failure must be small.

5. Record Linkage of MEPS Medical Events

Record linkage of MEPS medical events used five identifying fields: event dates (year, month, day, and day-of-week), medical condition codes, procedure codes, global-fee codes, and lengths (number of days) of hospital stay. These fields are described in more detail in Winglee, Valliant, Brick and Machlin (2000). A training sample from MEPS 1996 was employed to derive match rules and outcome categories and to estimate the probabilities of agreement for each category, allowing for partial agreement and value specific outcomes. The same match rules were repeated each year with minor adjustments of the matching parameters.

For the training set we used the linkage system Automatch (Matchware 1996) and the unique match algorithm to select linked pairs. In "unique" matching, a File A record is optimally linked to only one File B record (Jaro 1989). In addition, we used the many-to-many match algorithm to generate a random sample of nonlinked pairs to facilitate linkage error estimation. However, the methods for estimating error rates, described below, apply to any software that implements the linkage methods based on match weights. They are not specific to Automatch.

The tradeoff in determining the selection threshold for MEPS was between getting a high match rate and limiting mismatch linkage errors. A high threshold weight would minimize false positive (mismatch) errors at the expense of lowering the match rate and losing valuable data collected from medical providers. On the other hand, a low threshold

would increase false positive error and may affect the allocation of expenditure data in a way that would require special analytic techniques to overcome and even then only with uncertainty. Since both data sources had reported on ostensibly the same medical events for the same persons over the same period, the strategy was to maintain a reasonably high match rate and to conduct a manual review of a limited number of questionable linked pairs after selection to assess the analytic impact of falsely accepting them. Based on this decision the average match rate for the annual MEPS medical records files was about 85 percent.

The 1996 MEPS training sample M curve, labeled the "Tra- M " curve, was generated by applying match weights to "true" matched pairs for a random sample of 500 persons in MEPS 1996. For these persons, the manual review files contained 2,507 events from household respondents and 2,804 events from medical providers. Knowledgeable data managers reviewed the events and selected 1,501 pairs. We considered these as the true matched pairs in this evaluation. The manually matched pairs were assigned the weights derived from our match specification to generate a cumulative distribution function.

The 1996 training sample U curve, labeled the "Tra- U " curve, was generated using a random sample of unmatched pairs. We used a simple random sampling with replacement method to select 500 events each from the matching files and employed a many-to-many match algorithm to generate all 250,000 possible event pairs. For these randomly selected sets of pairs, the chance of there being any correctly matched pairs is negligible; thus, the entire set was taken to consist of unmatched pairs. We applied the match weights from our matching specification and plotted the "Tra- U " curve equal to 1 minus the cumulative distribution of the weights of these pairs. Figure 1 in section 8 shows both the Tra- M and Tra- U curves for the 1996 MEPS. The curves shown in this figure were smoothed using a nonparametric lowess function (Chamber, Cleveland, Kleiner and Tukey 1983) in S-PLUS 2000 (1999).

6. Simrate Implementation in MEPS

The SimRate weight distribution method used Monte Carlo simulation methods to generate separate sets of 10,000 simulated matched and unmatched pairs for creating the weight curves. To generate the "Sim- M " weight distributions we estimated the probabilities m_{vi} from linked pairs assigned by a unique matching algorithm. We used the "tuned" linkage system to select matched pairs from the 1996 annual matching files and tabulated the observed frequencies for each outcome category for each of the five matching fields. The proportion of pairs that fell into category i of field v was then used as the estimate \hat{m}_{vi} of the probability m_{vi} .

For the unmatched pairs and the “Sim-*U*” curve, the u_{vi} probabilities for unmatched pairs were estimated using the same sample of unmatched pairs used in creating the “Tra-*U*” curve. The difference is that we used these pairs to observe the relative frequencies for each outcome category for each of the five matching fields among unmatched pairs. The proportion of pairs that fell into category *i* of field *v* was then used as the estimate \hat{u}_{vi} of the probability u_{vi} .

For a simulated matched pair, a realization of the multinomial random variable y_{rv} was generated for each match field. For example, a configuration like (agreement on event date, agreement on length of hospital stay, agreement on the array of condition codes, joint agreement by type of procedure, and value specific agreement for a global-fee indicator) was generated using the match probabilities \hat{m}_{vi} for each outcome category. Similarly, for each unmatched pair, a realization was generated of a category for each of the five fields using the unmatched probabilities \hat{u}_{vi} discussed above.

For a given realization y_r , a weight w_r was computed for the pair by summing the weights for the randomly generated categories that the pair fell into. The actual weights used in our simulation were adjusted ones that we specified rather than ones defined directly by the matching software (see Winglee, *et al.* 2000). Thus, we are simulating the way in which matching would actually be implemented. To do this we calculated the match weight for both the matched and unmatched sets of 10,000 pairs and plotted the simulated match weight functions.

Table 2 shows examples of some the partial agreement categories used for matching event date and the estimates of \hat{m}_{vi} , \hat{u}_{vi} , and w_r used in SimRate simulation. We defined a total of 19 outcome categories for matching by event date, 9 categories for duration of hospital stay, 27 categories by medical procedures, and 3 categories each for medical conditions and global fee. For example, for the outcome category exact agreement on event date, the estimate of \hat{m}_{vi} was 0.69, meaning that 69 percent of the linked pairs had exact agreement on event date. The estimate of \hat{u}_{vi} for this outcome category was 0.003, showing that only 0.3 percent of the unlinked pair showed agreement on this field. The match weight for exact agreement on date of event was 8.52 and that for complete disagreement (difference of more than two weeks apart and on different day of week) was -6.64. (see Winglee, *et al.* 2000 for the match weights by match fields and outcome categories).

We selected the match fields that were approximately independent in this case study. For example, we found no functional association between the date of medical events and other match fields like medical condition and length of hospital stay. For fields such as the indicators for surgery, radiology, and laboratory procedures, we used chi-square

tests and found some dependence between the concurrence of surgery and radiology. To handle this situation, we estimated the joint probabilities and specified match rules to treat these procedure flags as a single match field (see section 4 above). Hence, we could then apply the independent multinomial distribution for simulation.

Table 2
Estimates of Multinomial Probabilities for Matched Pairs (\hat{m}_{vi}) and Unmatched Pairs (\hat{u}_{vi}), and Match Weights (w_{vi}) for the Match Field Event Date

Match rule for Event Date	\hat{m}_{vi}	\hat{u}_{vi}	w_{vi}
Missing	0.031	0.046	0.00
Exact match	0.693	0.003	8.52
Off +/- 1 day	0.068	0.006	5.71
Off +/- 3 day	0.023	0.005	4.09
Off +/- 5 day	0.014	0.005	2.47
Off +/- 7 day	0.030	0.006	2.84
Match by day of week only	0.014	0.034	-3.64
Disagree	0.003	0.547	-6.64

Table 3 shows the results of linkage error estimates from SimRate and the training curves at the threshold weight of $w=1$ for MEPS 1996, MEPS 1997, and MEPS 1998. SimRate was easy to repeat each year. Repeating the manual-based weight curves, however, depended in part on manual review and we had only one reliable training sample, that for 1996. Note that the linked pairs used in SimRate will naturally generate some percentage of false positives and false negatives, *i.e.*, some matched and unmatched pairs are incorrect. Thus, the \hat{m}_{vi} probabilities computed in this way for the identified fields are subject to error. It would have been preferable to estimate the *m* probabilities from a “truth” set where we were confident that all matches were correct. However, the manually matched training sets we were able to produce were too small to yield stable estimates in all of the detailed match categories and manual selection is also imperfect. This difference may explain in part the slightly higher overall error rate estimates from SimRate than from the training sample weight curves.

Table 3
Weight Curve Methods to Estimate Linkage Error Rates at Threshold Weight 1, MEPS 1996 – 1998

Method	Error Rate	1996	1997	1998
SimRate simulation curves	False negative	5.2	6.5	5.8
	False positive	9.0	6.9	7.6
Training sample curves	False negative*	3.3	3.3	3.3
	False positive**	5.5	6.4	5.7

* Estimates from the 1996 Tra-*M* curve were used for all three years.

** Estimates from the 1996 Tra-*U* curve used samples of 500 records from each match file and a total of 250,000 unmatched pairs. The 1997 and 1998 estimates used different Tra-*U* curves employing samples of 1,000 records from each match file and a total of 1,000,000 unmatched pairs.

7. Mixture Model Implementation in MEPS

A mixture modeling approach by Belin and Rubin (1995) views the distribution of observed match weights from a computerized linkage system as a mixture of weights for true matches and false matches. In principle, the mixture model method has two attractive features suitable for MEPS. First, it can handle repeated applications efficiently. When global parameter estimates of the transformed parameters and the ratio of the variances of the two distributions are available, these estimates can be applied to similar data for estimation. Since the MEPS record linkage is done annually, global estimates derived from early training samples could conceivably be applied for linkage error estimation in later years when manual review samples were not available.

The second advantage is that the mixture model can draw from multiple sets of parameter estimates from different training samples and can reflect variations. This feature is especially appealing for MEPS because manual review is a complex process and not necessarily always accurate. Hence, an alternative is to view the computer system selection as the truth and use them to provide an alternative set of parameter estimates. This process can also be repeated using training samples from more than one year.

Our application of the Belin–Rubin approach used the same training samples from MEPS 1996 and a second training sample of the same size from 1997. Following Belin–Rubin’s examples, we applied the mixture modeling method using manually identified true and false match pairs from a one-to-one matching system (note that such systems provide relatively few false match pairs for estimation). We computed model estimates for MEPS 1996 and MEPS 1997 assuming the manual selection to be the truth, and for testing the behavior of the model, we computed a second set of estimates assuming computer system selected match pairs to be the true pairs.

Implementation involved two procedures – the Box and Cox (1964) procedure for global parameter estimation and the Calibrate procedure (Belin and Rubin 1995) to fit a mixture model for error rate estimation. Before applying Box–Cox, the weights were rescaled between 1 and 1,000. The Box–Cox transformation discussed by Belin and Rubin (1995) was

$$\Psi(w_r) = \frac{w_r^\gamma - 1}{\gamma \bar{w}^{\gamma-1}}$$

where w_r is the match weight for pair r , \bar{w} is the geometric mean of the w_r weight, and γ is a parameter that is dependent on whether the pair is in the matched or unmatched set.

For the mixture model procedure to be effective, the transformed weights should be approximately normally distributed. The untransformed weight distribution with our data showed bimodality and almost no overlap in match weight between matched and unmatched pairs (bimodality was also observed in Belin–Rubin 1995). For example, application of their transformation procedure to the 1996 MEPS system pairs resulted in parameter estimates of $\bar{w} = 585.7$ and $\gamma = 1.15$ for the true matched pairs and $\bar{w} = 113.1$ and $\gamma = 0.48$ for the false matched pairs. The transformed weights, however, showed relatively little improvement towards normality. Since the match weights are the log of a product, or the sum of logs, we might hope that the weights would be normally distributed if there were many components in the sum. However, we had only five fields to use for matching. The small number of fields may have accounted in part for the lack of normality with our transformed data.

Table 4 shows the results of applying the Belin–Rubin mixture model to MEPS 1996. This table shows the model estimated false match rates, the 95 percent confidence interval of the estimated rate, and the actual observed false match rate at the threshold weight of 1. Using the manual review pairs as the true matched pairs, the model estimate of the expected false match rate at the threshold of $w = 1$ was 9.1 percent, with a 95 percent confidence interval ranging between 6.0 and 12.2. The actual observed false match error rate, however, was 14.5 percent, which is higher than the upper 95 percent confidence bound. Note that these are rates of the form $n_{12} / n_{1\bullet}$ in Table 1. These are not the same rates estimated by SimRate and the weight curve approach.

Table 4
Mixture Model Linkage Error Estimates

MEPS 1996	Percentage false match error			
	Expected rate	Lower Bound*	Upper Bound*	Observed rate
Manual match	9.1	6.0	12.2	14.5
System match	0.9	0.6	1.2	0.0

* The lower and upper bounds are the 95 percent confidence interval of the expected error rate.

Since manual review may not always be accurate, an option, for the purpose of evaluation, is to treat the computer system linked pairs as the truth matched pairs, and use them for modeling. Under this assumption, the model estimate of the expected error rate is 0.9, and a 95 percent confidence interval between 0.6 and 1.2. The actual observed rate in this case, 0 percent, was a hypothetical outcome treating the computer-linked pairs as correct. Of course, in reality there will be some nonzero level of error so that the mixture model confidence interval is not necessarily wrong.

We generated global parameter estimates using both the training sample manual selections and system selections for

MEPS 1996 and MEPS 1997 and used them as four sets of inputs to provide global estimates for modeling linkage error for MEPS 1998. This should be possible because the data remained similar and record pairs were selected using the same match rules for all 3 years. A difference was that manual review was not conducted for MEPS 1998 and we could not use the Box-Cox procedure for global parameter estimation for 1998 (because there was no separate manual indicator for true and false pairs). For this application, we use a bootstrap method in the Belin and Rubin Calibrate procedure to draw from multiple parameter sets to reflect uncertainties in estimation. This application, however, did not converge after 150 iterations of the estimation procedure. We could only conclude that the global parameter estimates from earlier training samples failed to generalize and provide error rate estimates for repeated linkage applications.

8. Concluding Comments and Analytic Implications

The process of threshold selection and linkage error estimation is an iterative process involving repeated cycles of observation, estimation, and modeling. Our case study

employed modeling approaches for estimating linkage errors and for monitoring the predictive power of the linkage system. Both methods provided valuable information for determining the linkage selection and for evaluating the quality of the declared matched pairs as we found in MEPS.

The weight curves approach of estimation has the appeal that one can choose a selection threshold to attain the acceptable linkage error level. For example, Figure 1 shows the training sample and the SimRate simulation weight curves based on the 1996 MEPS matching files. A vertical line is drawn at the selection threshold weight of $w = 1$; the error levels for 1996 MEPS (shown in Table 3) were then estimated by the cumulative percentage at threshold level. By sliding this threshold, one can aim to minimize the total linkage error by selecting a threshold at the crossing point of the M and U curves. In this case study, the optimal threshold suggested by both sets of weight curves is fairly consistent. We included a likelihood interpretation of the match weight. For example, at the match weight of $w = 1$, the likelihood ratio score is 2. This means that for records with a match weight of $w = 1$ or above, the relative likelihood of being true pairs is at least 2 to 1.

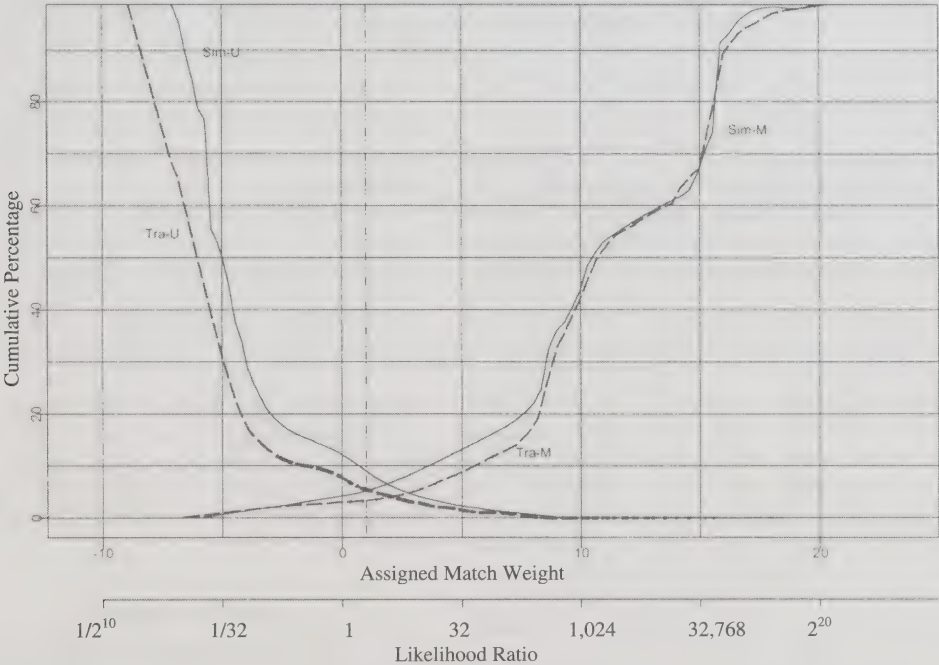


Figure 1. Weight Curves for MEPS 1996 using the SimRate and Training Sample Methods; the dashed vertical reference line shows the threshold value of 1.

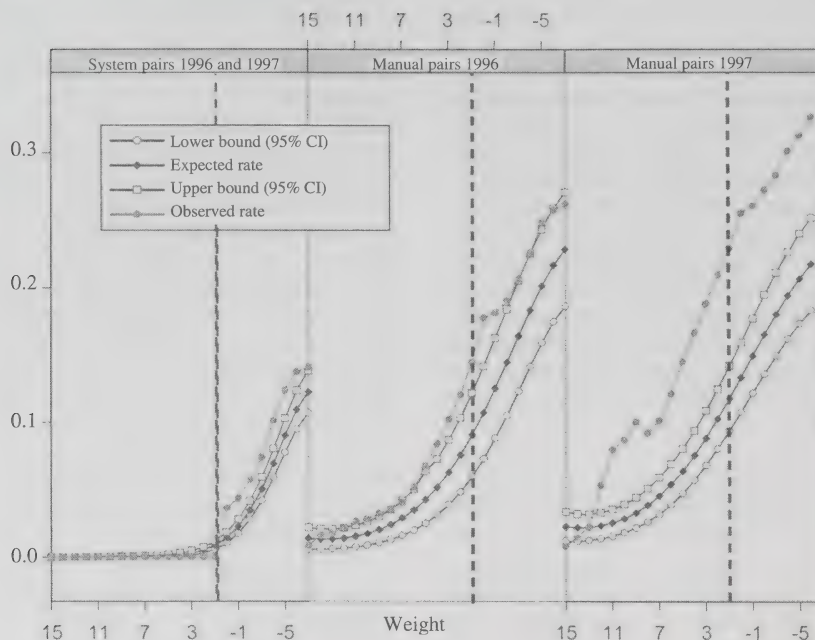


Figure 2. Mixture Model Estimates of False Match Rates by Weight, 1996 and 1997 MEPS Training Samples (a vertical line is drawn at weight = 1, which is threshold).

For linked pair quality, Figure 2 shows the distributions of false match rate estimates from mixture modeling. This figure shows the model estimated false match rate, the upper and lower 95 percent confidence bounds of the error rate estimates, and the actual observed rates. Panel 1 shows the estimates treating the computer system linked pairs as the true matched pairs. Panels 2 and 3 show the estimates from the 1996 MEPS and 1997 MEPS training samples. The difference between Panels 2 and 3 shows the inconsistency of manual selection by different reviewers in our application. In all three panels, the 95 percent confidence interval of the model estimates failed to cover the true observed values. Ideally, one would use both Figure 1 and Figure 2 together to guide the choice of selection thresholds.

We have found SimRate to be an informative and flexible tool for determining selection thresholds and estimating error rates in our application. Given multinomial or other models for the matching variables, the SimRate method provides error rate estimates that would be obtained from repeated application of the matching algorithm to a large number of candidate record pairs. It is also flexible in

accommodating the choices of comparison sets of pairs for computing rates.

While our application achieved the matching and error rate estimation goals for MEPS, more work might be done prior to or during the analysis stage. Space does not permit us to develop these in the context of the current case study but two general approaches might be mentioned. First, it is possible to reweight the final results and adjust for false nonmatches – treating them in a manner analogous to unit nonresponse (*e.g.*, as in Oh and Scheuren 1980). To handle mismatches, the ideas in Scheuren and Winkler (1993 and 1997), and Lahiri and Larsen (2002) might be worth consulting. Whether these added steps are needed, of course, depends on the final uses to which the linked data will be put.

Acknowledgements

The basic linkage research, reported on here, was conducted under contracts 290-99-0002 and 290-94-2002 sponsored by the Agency for Healthcare Research and

Quality and the National Center for Health Statistics. The authors would like to thank Steven B. Cohen, Steven Machlin, and Joel Cohen of the Agency for Healthcare Research and Quality for their comments on various stages of this research and Thomas Belin for his suggestions on an earlier draft.

References

- Agency for Healthcare Research and Quality (2001). MEP – Medical Expenditure Panel Survey. <<http://www.ahrq.gov/data/mepsix.htm>>.
- Armstrong, J.B., and Mayda, J.E. (1993). Model-based estimation of record linkage error rates. *Survey Methodology*, 19, 137-147.
- Bartlett, S., Krewski, D., Wang, Y. and Zielinski, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology*, 19, 3-12.
- Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations (with discussions). *Journal of the Royal Statistical Society, Series B*, 26, 206-252.
- Belin, T.R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Survey Methodology*, 19, 13-29.
- Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.
- Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P. (1983). *Graphic Methods for Data Analysis*, Duxbury Press, Boston.
- Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Fellegi, I.P. (1997). Record linkage and public policy – A Dynamic Evaluation. *Proceedings of the International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management and Budget*, Washington, DC.
- Gomatam, S., Carter, R., Ariet, A. and Mitchell, G. (2002). An empirical companion of record linkage procedures. *Statistics in Medicine*, 21, 1485-1496.
- Jaro, M.A. (1989). Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley & Sons, Inc.
- Lahiri, P., and Larsen, M.D. (2002). Regression analyses with linked data. (Draft manuscript).
- Larsen, M.D., and Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.
- Matchware Technologies Inc. (1996). *AutoMatch: Generalized Record Linkage System User's Manual*. Silver Spring, MD: Matchware Technologies, Inc.
- Newcombe, H.B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford University Press, New York.
- Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- Newcombe, H.B., and Kennedy, J.M. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Communications of the Association for Computing Machinery*, 5, 563-567.
- Oh, H.L., and Scheuren, F. (1980). Fiddling around with nonmatches and mismatches, *Studies from Interagency Data Linkages Series*. Social Security Administration, Report No. 11.
- Scheuren, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 377-381.
- Scheuren, F., and Winkler, W.E. (1993). Regression analyses of data files that are computer matched. *Survey Methodology*, 19, 35-58.
- Scheuren, F., and Winkler, W.E. (1997). Regression analyses of data files that are computer matched, II. *Survey Methodology*, 23, 157-165.
- S-Plus 2000 (1999). MathSoft, Inc. Data Analysis Products Division, Seattle, Washington.
- Tepping, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- Winglee, M., Valliant, R., Brick, J.M. and Machlin, S. (2000). Probability matching of medical events. *Journal of Economic and Social Measurement*, 26, 129-140.
- Winkler, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 829-834.
- Winkler, W.E. (1994). *Advanced Methods for Record Linkage*. Bureau of the Census Statistical Research Division, Statistical Research Report Series, RR 94/05.
- Winkler, W.E. (1995). *Matching and record linkage*. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College and P.S. Kott). New York: John Wiley & Sons, Inc., 355-384.

The Effect of Record Linkage Errors on Risk Estimates in Cohort Mortality Studies

D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski and R. Mallick¹

Abstract

The advent of computerized record linkage methodology has facilitated the conduct of cohort mortality studies in which exposure data in one database are electronically linked with mortality data from another database. This, however, introduces linkage errors due to mismatching an individual from one database with a different individual from the other database. In this article, the impact of linkage errors on estimates of epidemiological indicators of risk such as standardized mortality ratios and relative risk regression model parameters is explored. It is shown that the observed and expected number of deaths are affected in opposite direction and, as a result, these indicators can be subject to bias and additional variability in the presence of linkage errors.

Key Words: Cohort study; Computerized record linkage; Linkage errors; Linkage threshold weight; Poisson regression; Relative risk regression; Standardized mortality ratio.

1. Introduction

In recent years, a number of historical cohort studies have been carried out in environmental epidemiology using existing administrative databases as information sources (Howe and Spasoff 1986; Carpenter and Fair 1990). In general terms, this involves linking records of human exposure to environmental hazards with records on health status, often using computerized methods for matching individual records from different databases. In a cohort mortality study, the vital status of each cohort member is determined by linkage with mortality records maintained by government agencies. Excess mortality within the cohort relative to the general population may be due to exposures experienced by the cohort members.

In specific terms, record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same entity (Bartlett, Krewski, Wang and Zielinski 1993). Procedures for computerized record linkage (CRL) have become highly refined, using sophisticated algorithms to evaluate the likelihood of a correct match between two records (Hill 1988; Newcombe 1988). Statistics Canada has developed a CRL system called CANLINK which is capable of handling both single file linkages and linkages between two separate files (Howe and Lindsay 1981; Smith and Silins 1981). In this system, weights reflecting the likelihood of a match are attached to pairs of records. Two thresholds are set: potential matches

with linkage weights above the upper threshold are considered to be links whereas potential matches with weights below the lower threshold are considered to be nonlinks. Potential matches with weights between the upper and lower thresholds are resolved using additional information when available. Otherwise, a single threshold is selected to discriminate between links and nonlinks.

The confidentiality of records protected under the Statistics Act is strictly maintained in any study in which record linkage is employed. All studies requiring linkage with protected data bases must satisfy a rigorous review and approval process prior to implementation, following well-established procedures for data confidentiality (Singh, Feder, Duntzman and Yu 2001). All linked files with identifying information remain in the custody of Statistics Canada (Labossière 1986).

Computerized record linkage methods have been used to link environmental exposure data to the Canadian Mortality Data Base (CMDB). For example, a study of Canadian farm operators was initiated to investigate possible relationships between causes of death in over 326,000 farm operators in Canada and various socio-demographic and farming variables, particularly pesticide use (Jordan-Simpson, Fair and Poliquin 1990). In this study, the CMDB was linked with the 1971 Census of Population and the 1971 Census of Agriculture. Another ongoing large-scale study is based on the National Dose Registry (NDR) of Canada (Ashmore and Grogan 1985, Ashmore and Davies 1989). The NDR

1. D. Krewski, McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. To whom correspondence should be addressed; A. Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata, India; Y. Wang, Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2; S. Bartlett, Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2; J.M. Zielinski, Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2; R. Mallick, McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

contains information on occupational exposures to ionizing radiation experienced by over 400,000 Canadians dating back to 1950. The NDR has recently been linked to the CMDB to investigate associations between excess mortality due to cancer and occupational exposure to low levels of ionizing radiation (Ashmore, Krewski and Zielinski 1997; Ashmore, Krewski, Zielinski, Jiang, Semenciw and Létourneau 1998). More recently, the NDR has been linked to the Canadian Cancer Incidence Database (Sont, Zielinski, Ashmore, Jiang, Krewski, Fair, Band and Létourneau 2001). A comprehensive list of other health studies based on linking exposure data with the CMDB has been compiled by Fair (1989).

The success of record linkage studies depends on the quality of databases being linked (Roos, Soodeen and Jebamani 2001). Using population based longitudinal administrative data, Roos *et al.* examined data quality issues in studies of health and health care. Ardal and Ennis (2001) considered systematic errors in administrative databases involved in secondary analysis of health information. Although record linkage studies will benefit from the use of high quality data, limitations in data quality may be offset to a certain extent by the large sample sizes found in many administrative data bases.

Record linkage studies have several advantages over traditional epidemiological studies. By using existing administrative databases, the need to collect new data for health studies is circumvented, and large sample sizes can often be achieved with relatively little effort. Depending on the nature of the databases utilized, record linkage provides an inexpensive way of exploring many possible associations in epidemiological studies. Record linkage also has certain disadvantages. There is generally little control over the information collected, and there can be appreciable loss to follow-up. Another disadvantage of record linkage is the occurrence of linkage errors, which is the focus of this paper. Inevitably, some records that match will fail to be linked, and other nonmatching records will be incorrectly linked.

Relatively little work has been done to determine the impact of these linkage errors on statistical inferences. Neter, Maynes and Ramanathan (1965) used a simple linear regression model to analyze the impact of errors introduced during the matching process. Their results indicate that linkage errors inflate the residual variance and introduce bias into the estimated slope parameter. Winkler and Scheuren (1991) derived an expression for the bias in estimates of linear regression coefficients due to linkage errors. Advances in the estimation of linkage error rates by Belin and Rubin (1991) enabled Scheuren and Winkler (1993) to implement an improved bias adjustment procedure. Linear regression methods for the analysis of

computer matched data files are further discussed by Scheuren and Winkler (1997).

The purpose of this paper is to explore the impact of linkage errors on statistical inferences in cohort mortality studies. Relative risk regression models employed in the analysis of data from such studies are described in section 2, and expressions for the observed and expected numbers of deaths based on these models developed. The impact of linkage errors on the observed and expected number of deaths and person-years at risk is discussed in section 3. An analysis of the impact of linkage errors on estimates of standardized mortality ratios (SMRs) and relative risk regression parameters is given in section 4. Both types of errors can cause bias and additional variability in estimates of these parameters. Our conclusions are presented in section 5.

2. Relative Risk Regression Models

Statistical methods for the analysis of cohort mortality studies are well established (Breslow and Day 1987). The primary objective of such analysis is to determine if the exposure to the agent of interest increases the mortality rate among cohort members. Mortality is characterized by the hazard function, which specifies the death rate as a function of time. Letting T denote the time of death, the hazard function at time u is formally defined as

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{\Pr\{u \leq T < u + \Delta u | T \geq u\}}{\Delta u}. \quad (1)$$

Let $\lambda_i(u)$ denote the hazard function for a specific cause of death at time u for individual $i=1, \dots, N$ in a cohort of size N , and let $\mathbf{z}_i(u)$ represent a corresponding vector of covariates specific to that individual. We assume that the effect of these covariates is to modify the baseline hazard $\lambda^*(u)$ in accordance with the relative risk regression model

$$\lambda_i(u) = \lambda^*(u) \gamma\{\beta' \mathbf{z}_i(u)\}, \quad (2)$$

where γ is a positive function of the covariates and β is a vector of regression parameters.

Two special cases of the general relative risk regression model of particular interest are the multiplicative and additive risk regression models. Define the function γ in (2) by

$$\log \gamma(z) = \frac{(1+z)^\rho - 1}{\rho}. \quad (3)$$

When $\rho=1$, the general relative risk regression model reduces to be the multiplicative risk regression model

$$\lambda_i(u) = \lambda^*(u) \exp\{\beta' \mathbf{z}_i(u)\}, \quad (4)$$

This proportional hazards model was introduced by Cox (1972), and is widely used in the analysis of mortality data (Kalbfleisch and Prentice 1980). The additive risk regression model

$$\lambda_i(u) = \lambda^*(u) + \beta' z_i(u) \quad (5)$$

occurs as a limiting case as $\rho \rightarrow 0$.

Let t_i^0 and t_i^1 be the age at the time of entry into the study, and the age at the time of loss to follow-up (due to withdrawal from the study, termination of the study, or death) for the i^{th} subject of the cohort, respectively. Let $\delta_i = 1$ or 0, according to whether the i^{th} individual has or has not died at the time of loss to follow-up. The log-likelihood function based on the relative risk model (2) may be written as

$$\log L = \sum_{i=1}^N \left\{ \delta_i \log(\gamma\{\beta' z_i(t_i^1)\}) - \int_{t_i^0}^{t_i^1} \gamma\{\beta' z_i(u)\} \lambda^*(u) du \right\}. \quad (6)$$

When there is a single covariate $z_i(u) \equiv 1$, the maximum likelihood estimate of $\theta = \exp\{\beta\}$ reduces to the standardized mortality ratio $\text{SMR} = \text{OBS}/\text{EXP}$, where $\text{OBS} = \sum_{i=1}^N \delta_i$ and $\text{EXP} = \sum_{i=1}^N e_i$ are the observed and expected numbers of deaths, respectively, with $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u) du$.

Maximization of the likelihood function (6) can be computationally burdensome with large sample sizes. Breslow, Lubin and Langholz (1983) simplify the likelihood by assuming that the covariates take on constant values within states through which a subject passes during the course of the study. The states are defined by cross-classification of the covariates of interest. Specifically, suppose that there are J such states $\{S_j; j=1, \dots, J\}$ such that $z_i(u) = z_j$ whenever the i^{th} subject is in S_j at time u . These states are mutually exclusive and exhaustive, so that at any given time u , each member of the cohort will fall into one and only one state. The log-likelihood function (6) may then be written as

$$\log L = \sum_{j=1}^J \{d_{jj} \log(\gamma\{\beta' z_j\}) - \gamma\{\beta' z_j\} e_j\}, \quad (7)$$

where

$$e_j = \sum_{i=1}^N \int_{[z_i(u) \in S_j]} \lambda^*(u) du \quad (8)$$

is the contribution to the expected number of deaths from all person-years of observation in the state S_j , and d_{jj} denotes the total number of deaths in that state. Letting $\Lambda_j(\beta) = \log(\gamma\{\beta' z_j\})$, the maximum likelihood estimate $\hat{\beta}$ of β is obtained as the solution to the score equation

$$\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^J \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} \{d_{jj} - \exp\{\Lambda_j(\hat{\beta})\} e_j\} = 0. \quad (9)$$

3. The Effect of Linkage Errors on the Observed and Expected Numbers of Deaths

Two principal types of errors can occur when linking data files in CRL (Fellegi and Sunter 1969). A false positive occurs when a member of the cohort who is alive is incorrectly identified as dead, and a false negative occurs when a deceased member is considered to be alive. More specifically, for the mathematical development to follow, a false positive occurs in a particular state when an individual who remains alive throughout this state is incorrectly labelled as dead in this state. Similarly, a false negative occurs in a particular state when a member, who died before or during the sojourn in this state, is considered to be alive throughout this state. Within a particular state, false positives and false negatives thus represent special cases of misclassification error discussed by Anderson (1974, chapter 6.2.1). In this section, we will discuss the effect of these two types of linkage errors on the observed and expected numbers of deaths, respectively. To do this, we first define sets of indices within states which will be used to represent sets of correctly matched and incorrectly matched records.

3.1 Linkage Errors

Let A_j and D_j denote the set of labels for those individuals in the cohort who remain alive throughout state S_j , and those who are dead in S_j , respectively. Write D_{jj} as the subset of D_j corresponding to those individuals who have died in S_j . Let A_j^f , D_j^f and D_{jj}^f denote the corresponding sets in the presence of linkage errors. We further define D_j^p as the set of labels of those alive in S_j (that is, in A_j) but labeled as dead in S_j corresponding to the false positives in S_j . Similarly, A_j^N is the set of those dead in S_j (that is, in D_j) but labeled as alive in S_j corresponding to the false negatives in S_j . Let us also write D_{jj}^p as the subset of D_j^p corresponding to those who are labeled to have died in S_j and, similarly, A_{jj}^N as the subset of A_j^N who have died in S_j (that is, in D_{jj}). These sets satisfy the relations $A_j^f = (A_j - D_j^p) \cup A_j^N$, $D_j^f = (D_j - A_j^N) \cup D_j^p$, and $D_{jj}^f = (D_{jj} - A_{jj}^N) \cup D_{jj}^p$.

The effect of linkage errors on the likelihood function in (7) may be described as follows. Let t_{ij}^0 denote the time at which the i^{th} individual enters, actually or by linkage error, the j^{th} state S_j . Similarly, t_{ij}^1 denotes the time of death (if it occurs, actually or by linkage error) for the i^{th} individual in S_j and t_{ij}^2 the time of leaving S_j , actually or by linkage error. Note that, if t_{ij}^1 exists, it is less than or equal to t_{ij}^2 . Let us, for the sake of simplicity, assume that t_{ij}^1 , if it exists, is equal to t_{ij}^0 ; that is, all the deaths in a state occur at the corresponding entry times in that state. Although this will underestimate the expected number of deaths, for the

purpose of studying bias, it may not be that objectionable. Assuming all the deaths to occur at the times of leaving the corresponding states also offers similar simplification. Using (8) and the decomposition of A_j^L , the expected number of deaths e_j^L in S_j the presence of linkage errors can be written as

$$\begin{aligned} e_j^L &= \sum_{i \in A_j^L} \int_{t_0^i}^{t_1^i} \lambda^*(u) du \\ &= \sum_{i \in A_j} \int_{t_0^i}^{t_1^i} \lambda^*(u) du + \sum_{i \in A_j^D} \int_{t_0^i}^{t_1^i} \lambda^*(u) du \\ &\quad - \sum_{i \in D_j^L} \int_{t_0^i}^{t_1^i} \lambda^*(u) du \\ &= e_j - \Delta e_j, \end{aligned} \quad (10)$$

where

$$e_j = \sum_{i \in A_j} \int_{t_0^i}^{t_1^i} \lambda^*(u) du, \text{ and } \Delta e_j = e_j^P - e_j^N \quad (11)$$

with

$$e_j^P = \sum_{i \in D_j^P} \int_{t_0^i}^{t_1^i} \lambda^*(u) du \text{ and } e_j^N = \sum_{i \in A_j^N} \int_{t_0^i}^{t_1^i} \lambda^*(u) du. \quad (12)$$

For notational convenience, let us write $T_\lambda(i, j)$ for $\int_{t_0^i}^{t_1^i} \lambda^*(u) du$ in what follows. The term Δe_j represents the bias in the expected number of deaths in the j^{th} state due to linkage errors. It follows from (10) and (11) that the false positives tend to reduce the expected number of deaths and the false negatives tend to increase the expected number of deaths.

Using the decomposition for D_{jj}^L , the observed number of deaths d_{jj}^L in the presence of linkage errors may be written as

$$d_{jj}^L = d_{jj} + \Delta d_{jj}, \quad (13)$$

where

$$\Delta d_{jj} = d_{jj}^P - a_{jj}^N, \quad (14)$$

with d_{jj} , d_{jj}^P and a_{jj}^N denoting the number of individuals in the sets D_{jj} , D_{jj}^P and A_{jj}^N , respectively. The term Δd_{jj} represents the difference between the observed number of deaths in the j^{th} state due to linkage errors. It follows from (13) and (14) that the false positives will increase the observed number of deaths and the false negatives will reduce the observed number of deaths.

Vital status is often determined by linkage with the CMDB, which is generally much larger than the cohort of interest. When the exposure records of a live individual are incorrectly associated with those of a dead person, the deceased individual usually does not belong to the cohort. Thus, the person-years at risk contributed by the person remaining alive will end prematurely in the year of presumed death; the lost person-years at risk correspond to

the time period from the year of presumed death until the end of the follow-up. On the other hand, when the exposure records of a dead individual are incorrectly associated with those of a live person, the person-years at risk contributed by this individual will include an extra period from the actual death-year to the end of the follow-up. Thus, false positives will deflate the number of person-years at risk and false negatives will inflate the number of person-years at risk in the cohort.

3.2 Expectations and Variances of Differences Between the Observed and Expected Numbers of Deaths

The effect of linkage errors on the observed and expected numbers of deaths depends on the false positive and false negative rates. Let p_j^P and p_j^N denote the false positive and false negative rates, respectively, in S_j , for $j=1, \dots, J$, which are assumed to be constant within S_j and same for all the individuals in A_j and D_j , respectively. This assumption is reasonable whenever individuals in the same state are highly homogeneous, particularly with respect to attributes such as the quality of personal identifiers that influence linkage error rates. Although this idealized assumption is unlikely to be fully satisfied in practice, it affords considerable simplification in the subsequent evaluation of the effects of linkage errors. Formally, p_j^P (p_j^N) is the conditional probability that an individual in A_j (D_j) is labeled dead (alive) in S_j . That is, $p_j^P = P[i \in D_j^P | i \in A_j]$ and $p_j^N = P[i \in A_j^N | i \in D_j]$.

Let us write a_j , d_j , a_{jj}^N and d_{jj}^P as the number of individuals in A_j , D_j , A_{jj}^N and D_{jj}^P , respectively. Then, note that, d_j^P follows a *Binomial*(a_j , p_j^P) distribution and a_{jj}^N follows a *Binomial*(d_j , p_j^N) distribution. Also, d_{jj}^P follows a *Binomial*(a_j , p_{jj}^P) distribution, where p_{jj}^P is the conditional probability that an individual in A_j is labeled to have died in S_j . That is, $p_{jj}^P = P[i \in D_{jj}^P | i \in A_j]$. Clearly, $p_{jj}^P \leq p_j^P$. Similarly, a_{jj}^N follows a *Binomial*(d_{jj} , p_{jj}^N) distribution, where p_{jj}^N is the conditional probability that an individual in D_{jj} is labeled as alive in S_j . That is, $p_{jj}^N = P[i \in A_{jj}^N | i \in D_{jj}]$. Although there is no trivial relationship between p_{jj}^N and p_j^N in general, it is reasonable to assume $p_{jj}^N = p_j^N$ in this context of linkage errors.

Assuming that linkage errors related to different individuals are independent, the expectation and variance of the difference in the observed number of deaths in S_j , given by Δd_{jj} in (14), are

$$E[\Delta d_{jj}] = E[d_{jj}^P] - E[a_{jj}^N] = a_j p_{jj}^P - d_{jj} p_j^N \quad (15)$$

and

$$\begin{aligned} V[\Delta d_{jj}] &= V[d_{jj}^P] + V[a_{jj}^N] \\ &= a_j p_{jj}^P (1 - p_{jj}^P) + d_{jj} p_j^N (1 - p_j^N). \end{aligned} \quad (16)$$

Since A_j and D_{jj} consist of different sets of individuals, d_{jj}^P and a_{jj}^N are independent.

Similarly, the expectation and variance of the difference in the expected number of deaths in S_j , given by Δe_j in (11), can be calculated as follows. For this purpose, it is convenient to write e_j^P and e_j^N in terms of the following indicator variables. For $i \in A_j$, define $\xi_{ij} = I\{i \in D_j^P\}$ and $\xi_{ijj} = I\{i \in D_{jj}^P\}$. Also, for $i \in D_j$, define $\psi_{ij} = I\{i \in A_j^N\}$. Then, from (12) and the definitions of D_j^P and A_j^N , we have

$$e_j^P = \sum_{i \in A_j} \xi_{ij} T_\lambda(i, j) \quad (17)$$

and

$$e_j^N = \sum_{i \in D_j} \psi_{ij} T_\lambda(i, j). \quad (18)$$

In particular, one can write $d_{jj}^P = \sum_{i \in A_j} \xi_{ijj}$ and $a_{jj}^N = \sum_{i \in D_j} \psi_{ij}$, which are useful to derive (15) and (16). From (17) and (18), we have

$$\begin{aligned} E[\Delta e_j] &= E[e_j^P] - E[e_j^N] \\ &= p_j^P \sum_{i \in A_j} T_\lambda(i, j) - p_j^N \sum_{i \in D_j} T_\lambda(i, j), \end{aligned} \quad (19)$$

and

$$\begin{aligned} V[\Delta e_j] &= V[e_j^P] + V[e_j^N] \\ &= p_j^P (1 - p_j^P) \sum_{i \in A_j} T_\lambda^2(i, j) \\ &\quad + p_j^N (1 - p_j^N) \sum_{i \in D_j} T_\lambda^2(i, j), \end{aligned} \quad (20)$$

since A_j and D_j consist of different sets of individuals.

The results (15)–(16) and (19)–(20) indicate that record linkage errors will lead to bias and additional variation in the observed and expected number of deaths. Minimizing the variance terms in (16) and (20) is difficult since the two error rates p_j^P and p_j^N are not functionally independent. Generally, decreasing p_j^P will result in an increase in p_j^N and vice versa (see section 5 for further discussion of this point). Although these error rates are independent of the underlying relative risk regression model γ in (2), the mean square error obtained by combining the expectation and variance terms cannot be minimized without specification of the baseline hazard $\lambda^*(u)$, which appears in T_λ .

4. The Effect of Linkage Errors on Estimates of SMRs and Regression Coefficients

4.1 Standardized Mortality Ratios

To determine the effect of linkage errors on the SMR, we replace the actual observed and expected numbers of deaths

d_{jj} and e_j by the observed and expected number of deaths d_{jj}^P and e_j^P in the presence of linkage errors in the expression $\text{SMR} = \sum d_{jj} / \sum e_j$. Letting SMR_L denote the standardized mortality ratios in the presence of linkage errors, we have

$$\text{SMR}_L = \text{SMR} \left[1 + \frac{\sum \Delta d_{jj}}{\sum d_{jj}} \right] / \left[1 - \frac{\sum \Delta e_j}{\sum e_j} \right]. \quad (21)$$

It follows, from (10)–(14), that the false positives will increase the SMR, whereas the false negatives will decrease the SMR.

By using a first order Taylor series approximation of SMR_L about SMR , the difference $\Delta \text{SMR} = \text{SMR}_L - \text{SMR}$ can be expressed as

$$\frac{\Delta \text{SMR}}{\text{SMR}} = \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \quad (22)$$

Then, the mean and variance of the relative difference in the SMR can be approximated by

$$E\left[\frac{\Delta \text{SMR}}{\text{SMR}}\right] \approx \frac{\sum_j E[\Delta d_{jj}]}{\sum_j d_{jj}} + \frac{\sum_j E[\Delta e_j]}{\sum_j e_j} \quad (23)$$

and

$$\begin{aligned} V\left[\frac{\Delta \text{SMR}}{\text{SMR}}\right] &\approx \left(\sum_j d_{jj}\right)^{-2} V\left[\sum_j \Delta d_{jj}\right] \\ &\quad + \left(\sum_j e_j\right)^{-2} V\left[\sum_j \Delta e_j\right] \\ &\quad + 2 \left(\sum_j d_{jj}\right)^{-1} \left(\sum_j e_j\right)^{-1} \text{Cov}\left[\sum_j \Delta d_{jj}, \sum_j \Delta e_j\right], \end{aligned} \quad (24)$$

respectively. The right hand side of (23) can be easily calculated by using (15) and (19). In order to calculate the right hand side of (24), note that

$$\begin{aligned} V\left[\sum_j \Delta d_{jj}\right] &= \sum_j V[\Delta d_{jj}] \\ &\quad + 2 \sum_{j < j'} \text{Cov}[\Delta d_{jj}, \Delta d_{j'j'}], \end{aligned} \quad (25)$$

$$V\left[\sum_j \Delta e_j\right] = \sum_j V[\Delta e_j] + 2 \sum_{j < j'} \text{Cov}[\Delta e_j, \Delta e_{j'}], \quad (26)$$

and

$$\begin{aligned} \text{Cov}\left[\sum_j \Delta d_{jj}, \sum_j \Delta e_j\right] \\ = \sum_j \text{Cov}[\Delta d_{jj}, \Delta e_j] + \sum_{j \neq j'} \text{Cov}[\Delta d_{jj}, \Delta e_{j'}]. \end{aligned} \quad (27)$$

Without loss of generality, let us assume, for $j < j'$, that $t_{ij}^0 \leq t_{ij'}^0$ for the same individual i (alive or dead) in S_j and $S_{j'}$; that is, the entry time in S_j is the same or earlier than that in $S_{j'}$. We then have, for $j < j'$,

$$\text{Cov}[\Delta d_{jj}, \Delta d_{jj'}] = - \left(\sum_{i \in A_j \cap A_{j'}} p_{jj}^P p_{jj'}^P + \sum_{i \in A_j \cap D_{jj'}} p_{jj}^P p_{jj'}^N \right), \quad (28)$$

$$\begin{aligned} \text{Cov}[\Delta e_j, \Delta e_{j'}] &= \sum_{i \in A_j \cap A_{j'}} p_j^P (1 - p_{j'}^P) T_\lambda(i, j) T_\lambda(i, j') \\ &+ \sum_{i \in A_j \cap D_{j'}} p_j^P p_{j'}^N T_\lambda(i, j) T_\lambda(i, j') \\ &+ \sum_{i \in D_j \cap D_{j'}} p_j^N (1 - p_{j'}^N) T_\lambda(i, j) T_\lambda(i, j'), \quad (29) \end{aligned}$$

$$\begin{aligned} \text{Cov}[\Delta d_{jj}, \Delta e_j] &= \sum_{i \in A_j} p_{jj}^P (1 - p_j^P) T_\lambda(i, j) \\ &+ \sum_{i \in D_{jj}} p_j^N (1 - p_j^N) T_\lambda(i, j), \quad (30) \end{aligned}$$

$$\begin{aligned} \text{Cov}[\Delta d_{jj}, \Delta e_{j'}] &= \sum_{i \in A_j \cap A_{j'}} p_{jj}^P (1 - p_{j'}^P) T_\lambda(i, j') \\ &+ \sum_{i \in A_j \cap D_{j'}} p_{jj}^P p_{j'}^N T_\lambda(i, j') \\ &+ \sum_{i \in D_{jj} \cap D_{j'}} p_{j'}^N (1 - p_j^N) T_\lambda(i, j'), \text{ and} \quad (31) \end{aligned}$$

$$\begin{aligned} \text{Cov}[\Delta d_{jj'}, \Delta e_j] &= - \sum_{i \in A_j \cap A_{j'}} p_j^P p_{jj'}^P T_\lambda(i, j) \\ &+ \sum_{i \in A_j \cap D_{jj'}} p_j^P p_{j'}^N T_\lambda(i, j). \quad (32) \end{aligned}$$

Using (25) – (32), the variance of the relative difference $\Delta \text{SMR}/\text{SMR}$ can be approximated by the right hand side of (24). Two conclusions can be drawn from (23) and (24). First, linkage errors can lead to bias in the estimate of the SMR. Second, both types of linkage errors introduce additional variation into estimates of the SMR. Note that the first term in (32) is dominated by the first term in (29) for $p_j^P < 0.5$, and the negative covariance term (28) is dominated in the calculation of the variance in (25). Therefore, the additional variance (24) is strictly positive, since both the false positive and false negative rates are positive.

4.2 Relative Risk Regression Parameters

To determine the effect of linkage errors on regression parameter estimates, consider first the general relative risk regression model (2). Replacing the observed and expected numbers of deaths d_{jj} and e_j in the log-likelihood function

(7) with the observed and expected numbers of deaths in the presence of linkage errors d_{jj}^L and e_j^L , we have

$$\log L = \sum_{j=1}^J \{d_{jj}^L \log\{\gamma(\beta' \mathbf{z}_j)\} - \gamma(\beta' \mathbf{z}_j) e_j^L\}. \quad (33)$$

Let $\hat{\beta}$ and $\tilde{\beta}$ denote the maximum likelihood estimates of β based on $\{d_{jj}, e_j\}$ and $\{d_{jj}^L, e_j^L\}$, respectively. The score equation (9) can be written as

$$\sum_{j=1}^J \frac{\partial \Lambda_j(\tilde{\beta})}{\partial \beta} [d_{jj} + \Delta d_{jj} - \exp\{\Lambda_j(\tilde{\beta})\}(e_j - \Delta e_j)] = 0. \quad (34)$$

Assuming that $\Delta\beta = \tilde{\beta} - \hat{\beta}$ is small, a first order expansion of $\exp\{\Lambda_j(\tilde{\beta})\}$ around $\hat{\beta}$ gives

$$\exp\{\Lambda_j(\tilde{\beta})\} \approx \exp\{\hat{\Lambda}_j\} + \exp\{\hat{\Lambda}_j\} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \Delta\beta, \quad (35)$$

where $\hat{\Lambda}_j = \Lambda_j(\hat{\beta})$ and $\partial \hat{\Lambda}_j / \partial \beta$ is $\partial \Lambda_j / \partial \beta$ evaluated at $\beta = \hat{\beta}$. Substituting (35) into (34) leads to

$$\sum_{j=1}^J \frac{\partial \hat{\Lambda}_j}{\partial \beta} [d_{jj} - \exp\{\hat{\Lambda}_j\} e_j] + \sum_{j=1}^J \frac{\partial \hat{\Lambda}_j}{\partial \beta} \left[\Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j - \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \right] \approx 0. \quad (36)$$

Using (9), the first summation in (36) is zero. Consequently, since $\Delta e_j \Delta\beta$ is small, $\Delta\beta$ may be approximated by

$$\Delta\beta = \left(\sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \{\Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j\}. \quad (37)$$

It follows from (37) that

$$E[\Delta\beta] \approx \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \Lambda_j}{\partial \beta} \alpha_j, \quad (38)$$

where $\alpha_j = E[\Delta d_{jj}] + \gamma\{\hat{\beta}' \mathbf{z}_j\} E[\Delta e_j]$, which can be calculated from (15) and (19). Further,

$$\begin{aligned} V[\Delta\beta] &\approx \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \\ &\quad \left(\sum_j \sum_{j'} \frac{\partial \Lambda_j}{\partial \beta} \Theta_{jj'} \frac{\partial \Lambda_{j'}'}{\partial \beta} \right) \\ &\quad \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \quad (39) \end{aligned}$$

with

$\Theta_{jj'} = \text{Cov}[\Delta d_{jj} + \gamma\{\hat{\beta}'\mathbf{z}_j\}\Delta e_j, \Delta d_{jj'} + \gamma\{\hat{\beta}'\mathbf{z}_{j'}\}\Delta e_{j'}]$, which can also be easily obtained using (16), (20) and (28)–(32).

In the special case of the multiplicative risk model (4), the difference $\Delta\beta$ due to linkage errors may be approximated by

$$\Delta\beta = (X'WX)^{-1}X'(\Delta D + \Delta W), \quad (40)$$

where $X' = (\mathbf{z}'_1, \dots, \mathbf{z}'_J)$, $\Delta D' = (\Delta d_{11}, \dots, \Delta d_{JJ})$, $W = \text{diag}(\exp(\mathbf{z}'_1\hat{\beta})e_1, \dots, \exp(\mathbf{z}'_J\hat{\beta})e_J)$, and $\Delta W' = (\exp(\mathbf{z}'_1\hat{\beta})\Delta e_1, \dots, \exp(\mathbf{z}'_J\hat{\beta})\Delta e_J)$. Note that the weight matrix W is the Fisher information matrix for $\hat{\beta}$. It follows from (38) that

$$E[\Delta\beta] \approx (X'WX)^{-1}X'\Pi, \quad (41)$$

where $\Pi' = (\pi_1, \dots, \pi_J)$ with π_j being same as α_j , but $\gamma\{\hat{\beta}'\mathbf{z}_j\}$ replaced by $\exp(\mathbf{z}'_j\hat{\beta})$.

Further,

$$V[\Delta\beta] \approx (X'WX)^{-1}X'\Psi X(X'WX)^{-1}, \quad (42)$$

where Ψ is the matrix of $\Theta_{jj'}$'s with $\gamma\{\hat{\beta}'\mathbf{z}_j\}$ replaced by $\exp(\mathbf{z}'_j\hat{\beta})$. Note that (40)–(42) are special cases of (37)–(39), respectively, written in matrix notation.

With a single covariate $z_i = 1$, $X'WX = e^\beta \sum_j e_j$, $X'\Delta D = \sum_j d_{jj}$ and $X'\Delta W = e^\beta \sum_j \Delta e_j$. In this case,

$$\Delta\beta \approx \frac{\sum_j \Delta d_{jj} + e^\beta \sum_j \Delta e_j}{e^\beta \sum_j e_j}. \quad (43)$$

Since the $\text{SMR} = e^\beta = \sum_j d_{jj} / \sum_j e_j$, with $\Delta\beta = \Delta \text{SMR} / \text{SMR}$ in this case, we have

$$\Delta\beta = \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \quad (44)$$

Thus, (44) may be viewed as a special case of (22).

The preceding results indicate that both false positives and false negatives will introduce bias and additional variation into the estimates of relative risk regression parameters. The only negative contribution to this additional variance (39) is through $\text{Cov}[\Delta d_{jj}, \Delta d_{jj'}]$, given by (28), and the first term in (32) (see $\Theta_{jj'}$). Using the same argument as in section 4.1, it follows that this additional variance is strictly positive.

5. Conclusions

Record linkage is now a well-established technique in epidemiological studies of population health risks. By linking information on individual exposures from one database to information on health outcomes in another database, it is possible to construct large-scale informative

databases on risks to health of populations and population subgroups. The success of such studies will depend to a large extent on the quality of the two databases being linked, including the amount of information on individual identifiers used to link individuals in the two databases. In most studies, the accuracy of the linkage is examined by estimating the false link (false positive) and false nonlink (false negative) rates associated with the linkage process. In practice, this is usually done by drawing a sample of linked and nonlinked records, and determining the accuracy of the linkages in the sample using auxiliary information drawn from other sources.

Although CRL has been used for some time in cohort mortality studies, the impact of linkage errors on the reliability of statistical inferences drawn from such studies has not been subjected to detailed investigation. The theoretical results presented in this paper address this issue. These results show that in addition to inflating the observed number of deaths, false positives will tend to deflate the expected number of deaths. Conversely, false negatives inflate the expected numbers of deaths and deflate the observed number of deaths. Linkage errors were shown to introduce bias into estimates of SMRs. Relative risk regression coefficients are also subject to bias, the direction of which depends on the nature of the regression coefficient. In addition to these biases, linkage errors introduce additional uncertainty into estimates of both SMRs and regression coefficients.

Although we make the simplifying assumption of $t_{ij}^1 = t_{ij}^0$, one can derive the relevant expressions for bias and increased variability without this assumption; however, the expressions are too complex to offer additional insight into the effects of linkage errors. This is also true of the assumption that $p_{jj}^N = p_j^N$. There is a technical issue with the definition of A_j for the state(s) corresponding to the last age interval, which is usually open up to ∞ on the right hand side. In such state(s), the assumption that $t_{ij}^1 = t_{ij}^0$ will be problematic if the probability of dying in this last interval is appreciable. This problem may be circumvented by assuming the human life span to have a finite upper limit.

As discussed at the end of section 3.1, false positives occur primarily when an individual who is alive at the end of the follow-up period is incorrectly linked with a dead person. However, a person who died in one of the states S_j may be falsely linked with another person with an earlier death time. This leads to a false positive which persists until the actual time of death; the analysis in section 3 allows for this type of error. Similarly, a dead person may be falsely linked with another person dying at a later time, who is not alive at the end of follow-up. This case is treated as a false negative only up to the false death time. At this false time of death, this will contribute incorrectly to the number of

deaths, an error which has not been considered in section 3. However, this type of error would not normally be detected in typical record linkage studies in which a simplified manual check is used to identify false positives and false negatives. Since this type of error is likely to be rare, the effect is expected to be small.

In order to further explore the potential impact of linkage errors, let τ_j be the upper age limit for the j^{th} state S_j . (Note that some of the τ_j 's may be equal.) Then, letting α denote the probability of a linkage error (of either type), the false positive and negative rates, p_j^P and p_j^N , may be written as $\alpha P[T \leq \tau_j]$ and $\alpha P[T > \tau_j]$, respectively. In particular, $p_{jj}^P = \alpha P[\tau_{j-1} < T \leq \tau_j]$, where τ_{j-1} is the lower age limit for the j^{th} state, and $p_{jj}^N = p_j^N$. Therefore, the false positive rates may be greater than the false negative rates in the older age groups, with the reverse happening in the younger age groups. Assuming a similar pattern in the size of the D_j 's and A_j 's, some cancellation of terms may take place in the calculation of $E[\Delta e_j]$ in (19) and $E[\Delta d_{jj}]$ in (15). This cancellation effect will reduce the expected bias in the SMR and the relative risk regression parameters given in (23) and (38), respectively.

Although we have considered only all-cause mortality in this article, cause-specific mortality can be examined by simple modifications of the definitions of D_{jj} , D_{jj}^L and D_{jj}^P . These sets should then consider only those deaths from the specific cause of interest. Consequently, d_{jj} and e_j should denote, respectively, the observed and expected number of deaths of the specific type in S_j . The hazard function in (1) and (2) should relate to the specific type of death, with $\lambda^*(u)$ being the corresponding baseline cause-specific hazard rate. Finally, the indicator δ_i in section 2 should indicate the specific type of death.

While the preceding analytical results shed considerable light on the effects of linkage errors in cohort mortality studies, it is important to investigate such effects under conditions as close as possible as may be encountered in practice. To this end, we conducted a computer simulation study based on actual data from the National Dose Registry of Canada, in which the introduction of false links and false nonlinks with known probabilities have been used to further evaluate the impact of linkage errors on estimates of cancer risk (Mallick, Krewski, Dewanji and Zielinski 2002). These simulation results corroborate the theoretical findings of this paper.

While the results reported here may help to clarify the impact of linkage errors on statistical inference, methods that take such errors into account in the statistical analyses remain to be developed. Such methods may be based on response error models employed in survey sampling, used in conjunction with traditional statistical methods for analyses of cohort mortality data. Research in this area is underway.

6. Acknowledgements

This research was supported in part by a grant from the National Science and Engineering Research Council of Canada to D. Krewski, who currently holds the NSERC/SSHRC/McLaughlin Chair in Population Health Risk Assessment at the University of Ottawa. Preliminary versions of this paper were presented at the Annual Joint Meeting of the American Statistical Association in San Francisco, August 8-12, 1993, and the Annual Meeting of the Statistical Society of Canada, Montreal, July 10-16, 1995. The final draft was presented in the session in honour of J.N.K. Rao at the Statistics Canada Symposium 2001 held in Ottawa on October 18, 2001. The first author (D. Krewski) is particularly grateful to have been invited to speak in the session in honour of J.N.K. Rao, who served as his doctoral thesis supervisor many years ago. This work was completed while A. Dewanji was a Visiting Scholar at the McLaughlin Centre for Population Health Risk Assessment in the summer of 2002 and 2003.

References

- Anderson, T.W. (1974). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, Inc.
- Ardal, S., and Ennis, S. (2001). Data detectives: Uncovering systematic errors in administrative databases. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada, Ottawa.
- Ashmore, J.-P., and Grogan, D. (1985). The national dose registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.
- Ashmore, J.-P., and Davies, B.D. (1989). The national dose registry: A centralized record keeping system for radiation workers in Canada. In *Applications of Computer Technology to Radiation Protection*, IAEA-SR-136/58, J. Stephan Institute, Ljublyua, 505-520.
- Ashmore, J.-P., Krewski, D. and Zielinski, J.M. (1997). Protocol for a cohort mortality study of occupational radiation exposure based on the national dose registry of Canada. *European Journal of Cancer*, 33, S10-S21.
- Ashmore, J.-P., Krewski, D., Zielinski, J.M., Jiang, H., Semenciw, R. and Létoirneau, E. (1998). First analysis of occupational radiation mortality based on the national dose registry of Canada. *American Journal of Epidemiology*, 148, 564-574.
- Bartlett, S., Krewski, D., Wang, Y. and Zielinski, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology*, 19, 3-12.
- Belin, T.R., and Rubin, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.
- Breslow, N.E., Lubin, J.H. and Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78, 1-12.

- Breslow, N.E., and Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2: *The Design and Analysis of Cohort Studies*. IARC scientific publication No. 82, international agency for research on cancer, Lyon, France.
- Carpenter, M., and Fair, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference – 1989: Proceedings of Record Linkage Sessions & Workshop*. Ottawa Select Printing, Ottawa.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society*, B, 34, 187-220.
- Fair, M.E. (1989). Studies and References Relating to Uses of the Canadian Mortality Data Base. Report from the occupational and environmental health research unit, Health Division, Statistics Canada, Ottawa.
- Fellegi, I., and Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Hill, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy (Release 2.7). Report from research and general system, informatics services and development division, Statistics Canada, Ottawa.
- Howe, G.R., and Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.
- Howe, G.R., and Spasoff, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. University of Toronto Press, Toronto.
- Jordan-Simpson, D.A., Fair, M.E. and Poliquin, C. (1990). Canadian farm operator study: Methodology. *Health Reports*, 2, 141-155.
- Kalbfleish, J.D., and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.
- Labossière, G. (1986). Confidentiality and access to data: The practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, University of Toronto Press, Toronto.
- Mallick, R., Krewski, D., Dewanji, A. and Zielinski, J.M. (2002). A simulation study of the effect of record linkage errors in cohort mortality data. *Proceedings of International Conference in Recent Advances in Survey Sampling*. Carleton University, Ottawa, to appear.
- Neter, J., Maynes, E.S. and Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford Medical Publications. Oxford.
- Roos, L.L., Soodeen, R. and Jebamani, L. (2001). An information-rich environment: Linked-record systems and data quality in Canada. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada, Ottawa.
- Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.
- Scheuren, F., and Winkler, W.E. (1997). Regression analysis of data files that are computer matched—Part II. *Survey Methodology*, 23, 157-165.
- Singh, A.C., Feder, M., Dunteman, G. and Yu, F. (2001). Protecting confidentiality while preserving quality of public use micro data. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistics Canada, Ottawa.
- Smith, M.E., and Silins, J. (1981). Generalized iterative record linkage system. *Social Statistics Section, Proceedings of the American Statistical Association*, 128-137.
- Sont, W.N., Zielinski, J.M., Ashmore, J.P., Jiang, H., Krewski, D., Fair, M.E., Band, P. and Létourneau, E. (2001). First analysis of cancer incidence and occupational radiation exposure based on the national dose registry of Canada. *American Journal of Epidemiology*, 153, 309-318.
- Winkler, W.E., and Scheuren, F. (1991). How computer matching error effect regression analysis: Exploratory and confirmatory analysis. Technical report, Statistical research division, U.S. Bureau of the Census, Washington, D.C.

Analysis of Experiments Embedded in Complex Sampling Designs

Jan A. van den Brakel and Robbert H. Renssen¹

Abstract

At national statistical institutes, experiments embedded in ongoing sample surveys are conducted occasionally to investigate possible effects of alternative survey methodologies on estimates of finite population parameters. To test hypotheses about differences between sample estimates due to alternative survey implementations, a design-based theory is developed for the analysis of completely randomized designs or randomized block designs embedded in general complex sampling designs. For both experimental designs, design-based Wald statistics are derived for the Horvitz-Thompson estimator and the generalized regression estimator. The theory is illustrated with a simulation study.

Key Words: Design-based analysis; Measurement error models; Probability sampling; Randomized experiments; Superimposition.

1. Introduction

A part of survey methodology is to consider and test alternative survey methods, to improve the quality and efficiency of sample survey processes at national statistical institutes. Large-scale field experiments embedded in ongoing surveys are particularly appropriate to quantify the effect of alternative survey implementations on response behavior or estimates of finite population parameters. At Statistics Netherlands, for example, the effects of alternative questionnaire designs, different approach strategies or different advance letters have been investigated on both kinds of parameters, see Van den Brakel and Renssen (1998), Van den Brakel (2001), and Van den Brakel and Van Berkel (2002). At national statistical institutes, sample surveys are generally kept unchanged as long as possible in order to construct uninterrupted time series of estimates of population parameters. It is inevitable, however, that survey processes are adjusted from time to time. Embedded experiments can be applied to detect and quantify possible trend disruptions in these time series due to necessary changes to a sample survey and provide a safe transition from an old to a new survey design. Running the old and new surveys concurrently by means of an embedded experiment creates the possibility of falling back on the old approach for regular publication purposes if the new approach turns out to be a failure.

Applications of embedded experiments in the literature are aimed at the estimation of the bias or the various variance components in total measurement error models. Mahalanobis (1946) introduced the idea of embedding experiments in ongoing sample surveys, probably for the first time, as interpenetrating subsampling to test interviewer differences under simple random sampling and unrestricted

randomization of sampling units to interviewers. Fellegi (1964) and Hartley and Rao (1978) generalized this approach to estimate response variances under more complex sampling designs and restricted randomization of sampling units. Fienberg and Tanur (1987, 1988, 1989) discuss the differences and parallels between the theory of experimental designs and finite population sampling and how the statistical methodology employed in both fields can be combined in a useful and natural way in the design and analysis of embedded experiments. In their 1988 article, they give a comprehensive overview of applications of embedded experiments mentioned in the literature.

The typical situation considered in this paper is a field experiment designed to compare the effect of K different survey implementations, *i.e.*, the treatments, on the main estimates of the finite population parameters of a current survey. To this end, a probability sample that is drawn from a finite target population is randomly divided into K subsamples according to an experimental design. Each subsample is assigned to one of the K treatments. The experimental designs considered in this paper are completely randomized designs (CRD's) and randomized block designs (RBD's) where sampling structures like strata, primary sampling units (PSU's), clusters or interviewers are potential block variables. Generally one large subsample is assigned to the regular survey, which will be used for official publication purposes and which will simultaneously serve as the control group in the experiment. The purpose of embedded experiments is the estimation of finite population parameters under the different survey implementations and to test hypotheses about the differences between estimates of those parameters.

At first instance, a standard model-based approach might be considered for this analysis. Since experimental units are

1. Jan A. van den Brakel and Robbert H. Renssen, Statistics Netherlands, Department of Statistical Methods, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands.

drawn by means of a complex sampling design without replacement from a finite population, the application of such an approach might result in design-biased parameter and variance estimates. This makes the analysis results incommensurate with the parameter and variance estimates of the ongoing survey, which complicates the interpretation of the results in the design-based setting of the sample survey. To make the analysis more robust to departures from the assumed model, a design-based analysis that accounts for the sampling design should be applied.

Before we present our design-based approach two alternatives are mentioned that, at first glance, seem to be correct. We briefly argue, however, that both alternatives generally give invalid results. The first alternative is to apply a design-based linear regression analysis that accounts for the sampling design to estimate and test hypotheses about the K treatment effects in the regression model. This approach easily results, however, in wrong design variances, since the randomization of the experimental design is ignored. The main analysis objective of embedded experiments is to compare the effect of alternative survey approaches on the main estimates of the current sample survey. A linear regression analysis doesn't precisely meet this objective, since the treatment effects in the regression model are generally not equal to the differences between the subsample estimates.

The second alternative is to apply a design-based inference for comparing domain parameters, in which the K treatments are considered as K domains. The objective of an embedded experiment, however, is to compare estimates of the same parameter under different survey strategies or treatments, whereas in the case of domain parameters the objective is to compare estimates of different population parameters under basically the same survey strategy.

The approach presented in this paper can be summarized as follows. Based on the K subsamples, a design-based estimator for the population parameter observed under each of the K treatments, and a design-based estimator for the covariance matrix of the $K - 1$ contrasts between these estimates are derived. This estimation procedure accounts for the probability structure of the sampling design, the random assignment of sampling units to treatments due to the experimental design, and the weighting procedure applied in the ongoing survey for the estimation of target parameters. This gives rise to a design-based Wald statistic to test the stated hypotheses about differences between sample survey estimates.

The main contribution of this paper is to provide a general framework for comparing K alternative survey approaches in the realistic situation of a full-scale sample survey process. The random selection of sampling units from a finite target population by means of a probability

sample is used in combination with randomization of the sampling units over different treatments according to an experimental design. This facilitates comparison of alternative survey implementations on the main outcomes of a sample survey and the generalization of the observed results to populations larger than the sample included in the experiment. The analysis procedure proposed in this paper generalizes the analysis of two-treatment experiments embedded in sample surveys (Van den Brakel and Renssen (1998) and Van den Brakel and Van Berkel (2002)) to CRD's and RBD's with $K > 2$ treatments. An important result is that the design-based estimator for the covariance matrix of the contrasts between the subsample estimates has a relatively simple structure, as if the sampling units were drawn with replacement and unequal selection probabilities. As a result neither joint inclusion probabilities nor design-covariances between the subsample estimates are required in the variance estimation procedure. This results in an attractive and relatively simple analysis procedure. A second advantage is that this procedure tests hypotheses on differences between the sample estimates of the survey, which facilitates the interpretation of the analysis results in many applications.

A design-based theory for the analysis of embedded experiments is presented in section 2. In section 3 it is explained in more detail why the design-based linear regression analysis is less appropriate. In section 4, the proposed design-based analysis procedure is evaluated in a simulation study. Conclusions are summarized in section 5.

2. Analysis of Embedded Experiments

2.1 Measurement Error Models

Although the analysis procedure for embedded experiments proposed in this section is design-based, some use is made of measurement error models. Testing systematic effects of different survey methodologies on the outcomes of a survey implies the existence of measurement errors. The traditional notion that observations obtained from sampling units are true fixed values observed without error, generally assumed in design-based sampling theory, is not tenable in such situations. Therefore a measurement error model is specified for the observations obtained under the different survey implementations or treatments of the experiment. This model links the treatment effects to systematic differences between finite population parameters.

Consider a finite population U of N individuals. Let variable y_{ikl} denote the potential response of the i^{th} individual ($i = 1, 2, \dots, N$) observed by means of the k^{th} treatment ($k = 1, 2, \dots, K$) and the l^{th} interviewer ($l = 1, 2, \dots, L$). It is assumed that these observations are a

realization of the measurement error model $y_{ikl} = u_i + \beta_k + \psi_{il} + \varepsilon_{ik}$. Here u_i is the true, intrinsic value of the i^{th} individual, β_k the effect of the k^{th} treatment, ψ_{il} the effect of the l^{th} interviewer on the i^{th} individual and ε_{ik} an error component of the i^{th} individual observed by means of the k^{th} treatment. The interviewer effect ψ_{il} allows for systematic clustering and correlation between the responses of the individuals assigned to the same interviewer due to fixed and random interviewer effects, *i.e.*, $\psi_{il} = \psi_l + \xi_{il}$, with ψ_l the fixed and ξ_{il} the random effect of the l^{th} interviewer. Besides interviewers, common factors such as coders and supervisors might also induce correlation between the responses of the individuals.

Since for each sampling unit a potential response variable is defined for each of the K different treatments, the measurement error model can be expressed in matrix notation as

$$\mathbf{y}_{il} = \mathbf{j}u_i + \boldsymbol{\beta} + \mathbf{j}\psi_{il} + \boldsymbol{\varepsilon}_i, \quad (1)$$

where $\mathbf{y}_{il} = (y_{il1}, \dots, y_{ilK})^t$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^t$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iK})^t$ and $\mathbf{j} = (1, \dots, 1)^t$. Let E_m and Cov_m denote the expectation and the covariance with respect to the measurement error model. The following model assumptions are made:

$$E_m(\boldsymbol{\varepsilon}_i) = \mathbf{0}, \quad (2)$$

$$\text{Cov}_m(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}^t) = \begin{cases} \sum_i : & i = i' \\ \mathbf{0} : & i \neq i' \end{cases}, \quad (3)$$

$$E_m(\xi_{il}) = 0, \quad (4)$$

$$\text{Cov}_m(\xi_{il}, \xi_{i'l'}^t) = \begin{cases} \tau_l^2 : & l = l' \\ 0 : & l \neq l' \end{cases}, \quad (5)$$

$$\text{Cov}_m(\varepsilon_{ik}, \xi_{il}) = 0, \quad (6)$$

where $\mathbf{0}$ is a vector of order K with each element zero and \mathbf{O} a matrix of order $K \times K$ with each element zero. If $\psi_l = 0$, then a model with only random interviewer effects is obtained. If $\tau_l^2 = 0$, then a model with only fixed interviewer effects is obtained. From the assumptions, it follows that

$$E_m(\mathbf{y}_{il}) = \mathbf{j}u_i + \mathbf{j}\psi_l + \boldsymbol{\beta}, \quad (7)$$

and

$$\text{Cov}_m(\mathbf{y}_{il}, \mathbf{y}_{i'l'}^t) = \begin{cases} \sum_i + \mathbf{j}\mathbf{j}^t \tau_l^2 : & i = i' \text{ and } l = l' \\ \mathbf{j}\mathbf{j}^t \tau_l^2 : & i \neq i' \text{ and } l = l' \\ \mathbf{O} : & i \neq i' \text{ and } l \neq l' \end{cases} \quad (8)$$

Any correlation between the responses of different individuals can be modeled by means of random interviewer effects. Any fixed interviewer effects influence the expected

response values. From now on, for notational convenience, the subscript l will be omitted in y_{ikl} and \mathbf{y}_{il} .

2.2 Hypotheses Testing

The measurement error model for the observations obtained in the experiment enables us to relate systematic differences between population parameters to the different survey implementations. Suppose that L interviewers are available for the data collection. The population U of size N can conceptually be divided into L groups U_l of size N_l , $l = 1, \dots, L$, such that all individuals within a group are potentially interviewed by the same interviewer. Let $\bar{\mathbf{Y}} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_K)^t$ denote the K dimensional vector of population means of \mathbf{y}_i , *i.e.*,

$$\bar{\mathbf{Y}} = \mathbf{j} \frac{1}{N} \sum_{i=1}^N u_i + \boldsymbol{\beta} + \mathbf{j} \sum_{l=1}^L \frac{N_l}{N} \psi_l + \mathbf{j} \sum_{l=1}^L \frac{N_l}{N} \xi_l + \frac{1}{N} \sum_{i=1}^N \boldsymbol{\varepsilon}_i. \quad (9)$$

The objective of the experiment is to investigate whether there are systematic differences between the K population means of $\bar{\mathbf{Y}}$ due to the K different survey strategies or treatments. This can be accomplished by formulating hypotheses about

$$E_m(\bar{\mathbf{Y}}) = \mathbf{j} \frac{1}{N} \sum_{i=1}^N u_i + \mathbf{j} \sum_{l=1}^L \frac{N_l}{N} \psi_l + \boldsymbol{\beta}, \quad (10)$$

where the expectation is taken over the measurement error model. This gives rise to the following hypothesis:

$$\begin{aligned} H_0 : \mathbf{C} E_m \bar{\mathbf{Y}} &= \mathbf{0}, \\ H_1 : \mathbf{C} E_m \bar{\mathbf{Y}} &\neq \mathbf{0}, \end{aligned} \quad (11)$$

where \mathbf{C} denotes a $(K-1) \times K$ matrix with $K-1$ contrasts and $\mathbf{0}$ a $K-1$ vector of zeros. Since $\mathbf{C}\mathbf{j} = \mathbf{0}$, it follows that $\mathbf{C} E_m \bar{\mathbf{Y}} = \mathbf{C}\boldsymbol{\beta}$ and hypothesis (11) concerns the treatment effects as represented by $\boldsymbol{\beta}$ in the measurement error model (1). The contrasts between the population parameters neatly correspond to these treatment effects. For the randomized experiments considered in this paper, it holds that each experimental unit assigned to an interviewer l has a nonzero probability of being assigned to each of the K treatments. Therefore, the bias in the parameter estimates due to fixed interviewer effects is the same under each of the K treatments and cancels out in the $K-1$ contrasts between the K parameter estimates.

Hypothesis (11) will be tested by estimating $E_m \bar{\mathbf{Y}}$ instead of $\boldsymbol{\beta}$, taking into account the sampling design, the experimental design, and the weighting procedure of the ongoing survey applied for the estimation of population parameters. To test (11), a probability sample drawn from a finite population is available. The sampling units (experimental units) are randomized over K subsamples and are assigned to one of the K treatments. In section 2.3 a

design-unbiased estimator for $E_m \bar{\mathbf{Y}}$, denoted $\hat{\bar{\mathbf{Y}}}$ is derived. For example $\hat{\bar{\mathbf{Y}}}$ may be the Horvitz-Thompson estimator or the generalized regression estimator. Let \mathbf{V} denote the covariance matrix of $\hat{\bar{\mathbf{Y}}}$. An (approximately) design-unbiased estimator for the covariance matrix of the $K-1$ contrasts of $\hat{\bar{\mathbf{Y}}}$, denoted $\hat{\mathbf{C}}\hat{\mathbf{V}}\hat{\mathbf{C}}'$, will be derived in section 2.4. Now, hypothesis (11) can be tested by means of the following design-based Wald statistic:

$$W = \hat{\bar{\mathbf{Y}}}^t \mathbf{C}^t (\hat{\mathbf{C}}\hat{\mathbf{V}}\hat{\mathbf{C}}^t)^{-1} \hat{\mathbf{C}}\hat{\bar{\mathbf{Y}}}. \quad (12)$$

For mathematical convenience, we prefer the contrast matrix $\mathbf{C} = (\mathbf{j}; -\mathbf{I})$, where \mathbf{j} is a $K-1$ vector of ones and \mathbf{I} the $(K-1) \times (K-1)$ identity matrix.

2.3 Estimation of Treatment Effects

2.3.1 Horvitz-Thompson Estimator

Consider a sample s drawn by a generally complex sampling design, that can be described by the first and second order inclusion probabilities π_i and π_{it} of the i^{th} and i, t^{th} sampling unit(s) respectively. In the case of a CRD, sample s is randomly divided into K subsamples s_k of size n_k . If $n_+ = \sum_{k=1}^K n_k$ denotes the number of sampling units in s , then the conditional probability that the i^{th} sampling unit is selected in subsample s_k , given that sample s is selected, is equal to n_k / n_+ . In the case of an RBD the sampling units are, conditionally on the realization of s , deterministically divided into J blocks s_j . Potential block variables are sampling structures like strata, clusters, PSU's, interviewers and the like. Within each block, the sampling units are randomized over the K treatments. Let n_{jk} denote the number of sampling units in block j assigned to treatment k . Then $n_{j+} = \sum_{k=1}^K n_{jk}$ denotes the size of block j , $n_{+k} = \sum_{j=1}^J n_{jk}$ denotes the size of subsample s_k and $n_{++} = \sum_{k=1}^K \sum_{j=1}^J n_{jk}$ denotes the size of sample s . The conditional probability that the i^{th} sampling unit is selected in subsample s_k , given that sample s is selected and $i \in s_j$, is equal to n_{jk} / n_{j+} .

Each subsample s_k can be considered as a two-phase sample, where the first order inclusion probabilities of the first phase sample are obtained from the sampling design and the conditional first order inclusion probabilities of the second phase sample are obtained from the experimental design. From this point of view, the first order inclusion probabilities for the elements of s_k are equal to $\pi_i^* = (n_k / n_+) \pi_i$ for CRD's and $\pi_i^* = (n_{jk} / n_{j+}) \pi_i$ for RBD's if this i^{th} sampling unit is assigned to the j^{th} block. It follows that the Horvitz-Thompson estimator for \bar{Y}_k , based on the n_{+k} observations obtained from subsample s_k can be defined as:

$$\hat{Y}_{k, \text{HT}} = \frac{1}{N} \sum_{i=1}^{n_{+k}} \frac{y_{ik}}{\pi_i^*} \equiv \frac{1}{N} \sum_{i=1}^{n_{+k}} \frac{\mathbf{p}_{ik}^t \mathbf{y}_i}{\pi_i}, \quad (13)$$

where \mathbf{p}_{ik} are K -vectors that describe the randomization mechanism of the experimental design. For a CRD, it follows that

$$\mathbf{p}_{ik} \equiv \begin{cases} \frac{n_+}{n_k} \mathbf{r}_k & \text{if } i \in s_k, \\ \mathbf{0} & \text{if } i \notin s_k \end{cases}, \quad (14)$$

and for an RBD

$$\mathbf{p}_{ik} \equiv \begin{cases} \frac{n_{j+}}{n_{jk}} \mathbf{r}_k & \text{if } i \in s_{jk}, \\ \mathbf{0} & \text{if } i \notin s_{jk} \end{cases}, \quad (15)$$

where \mathbf{r}_k denotes the unit vector of order K with the k^{th} element equal to one and the other elements equal to zero and $\mathbf{0}$ denotes a K vector of zeros. Properties of the vectors \mathbf{p}_{ik} are given in the appendix.

Now, since s_k can be considered as a two-phase sample it holds that $E_s E_e (\hat{Y}_{k, \text{HT}} | s, m) = \bar{Y}_k$, where E_s and E_e denote the expectation with respect to the sample design and the experimental design, respectively. So, given m , the vector $\hat{\bar{\mathbf{Y}}}_{\text{HT}} = (\hat{Y}_{1, \text{HT}}, \dots, \hat{Y}_{K, \text{HT}})^t$ is proposed as a design-unbiased estimator for $\bar{\mathbf{Y}}$. But then, $\hat{\bar{\mathbf{Y}}}_{\text{HT}}$ is unbiased for $E_m \bar{\mathbf{Y}}$.

2.3.2 The Generalized Regression Estimator

In finite population sampling it is customary to increase the accuracy of the Horvitz-Thompson estimator, if suitable auxiliary information is available, by means of the generalized regression estimator, see *e.g.*, Bethlehem and Keller (1987) and Särndal, Swensson and Wretman (1992). The generalized regression estimator enables us to incorporate the weighting scheme of the ongoing survey in the analysis of embedded experiments. This might decrease the design variance as well as the bias due to selective nonresponse and therefore it may increase the accuracy of the experiment. In the present context the generalized regression estimator therefore represents a design-based analogue of covariance analysis in standard experimental design methodology.

Besides the values of the response variable y_i , we also associate with each unit in the population an H -vector \mathbf{x}_i , of auxiliary information. The finite population means of these auxiliary variables are assumed to be known and are denoted by $\bar{\mathbf{X}}$. It is also assumed that the auxiliary variables are intrinsic values, that can be observed without measurement errors, and so are not affected by the treatments. When the model assisted approach of Särndal *et al.* (1992) is followed, the intrinsic values u_i in the measurement error model of section 2.1 for each unit in the population are assumed to be an independent realization of the following linear regression model:

$$u_i = B'x_i + e_i, \quad (16)$$

where B is an H -vectors containing the regression coefficients and the e_i are the residuals. In the model assisted approach of Särndal *et al.* (1992), the intrinsic values u_i are considered to be a realization of an underlying superpopulation model defined by (16). In this case the residuals e_i are independent random variables with a variance ω_i^2 . Then it is required that all ω_i^2 are known up to a common scale factor; that is $\omega_i^2 = v_i \omega^2$ with v_i known. From a strictly design-based point of view, proposed by Bethlehem and Keller (1987), there is no need to adopt a superpopulation model. In that case the residuals are fixed intrinsic values of the elements in the finite population and no model assumptions about the residuals are needed. In this paper, the model assisted approach of Särndal is adopted. This implies that expectations with respect to the measurement model, as in (7) and (10), are conditional on the realization of the intrinsic values $u_i, i=1, \dots, N$, in the finite population according to the superpopulation model (16).

The regression coefficients of the linear model (16) in the finite population are defined as

$$\mathbf{b} = \left(\sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i=1}^N \frac{\mathbf{x}_i u_i}{\omega_i^2}. \quad (17)$$

The intrinsic values u_i are not observable due to measurement errors and treatment effects. Consequently, (17) cannot be computed, even in the case of a complete enumeration of the finite population. In the case of a complete enumeration under the k^{th} treatment

$$\tilde{\mathbf{b}}_k = \left(\sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i=1}^N \frac{\mathbf{x}_i y_{ik}}{\omega_i^2}, \quad k = 1, 2, \dots, K, \quad (18)$$

denotes the finite population regression coefficients of the linear model (16). Conditional on the realization of $u_i, i=1, \dots, N$, the expectation of the finite population regression coefficients $\tilde{\mathbf{b}}_k$ with respect to the measurement error model is given by

$$E_m \tilde{\mathbf{b}}_k = \left(\sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i=1}^N \frac{\mathbf{x}_i (u_i + \beta_k + \psi_i)}{\omega_i^2} \equiv \mathbf{b}_k, \quad k = 1, 2, \dots, K. \quad (19)$$

The finite population regression coefficients $\tilde{\mathbf{b}}_k$ and \mathbf{b}_k can be estimated using the sample data from subsample s_k , with the Horvitz-Thompson estimator:

$$\hat{\mathbf{b}}_k = \left(\sum_{i=1}^{n_{k,s}} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i} \right)^{-1} \sum_{i=1}^{n_{k,s}} \frac{\mathbf{x}_i y_{ik}}{\omega_i^2 \pi_i}, \quad k = 1, 2, \dots, K.$$

Now the generalized regression estimator for \bar{Y}_k , based on the $n_{k,s}$ observations of subsample s_k , is defined as

$$\hat{\bar{Y}}_{k:\text{greg}} = \hat{\bar{Y}}_{k:\text{HT}} + \hat{\mathbf{b}}_k' (\bar{\mathbf{X}} - \hat{\mathbf{X}}_{\text{HT}}), \quad k = 1, 2, \dots, K, \quad (20)$$

where $\hat{\mathbf{X}}_{\text{HT}}$ denotes the Horvitz-Thompson estimator for the population means of the auxiliary variables $\bar{\mathbf{X}}$ based on the $n_{k,s}$ sampling units of subsample s_k .

When expressing (20) as a function of $(\hat{\bar{Y}}_{k:\text{HT}}, \hat{\mathbf{b}}_k, \hat{\mathbf{X}}_{\text{HT}})$, the generalized regression estimator can be approximated by means of a first order Taylor linearization about $(E_m \bar{Y}_k, \mathbf{b}_k, \bar{\mathbf{X}})$, where \mathbf{b}_k is defined in (19). This gives:

$$\hat{\bar{Y}}_{k:\text{greg}} \doteq \hat{\bar{Y}}_{k:\text{HT}} + \mathbf{b}_k' (\bar{\mathbf{X}} - \hat{\mathbf{X}}_{\text{HT}}) = \hat{\bar{E}}_{k:\text{HT}} + \mathbf{b}_k' \bar{\mathbf{X}}, \quad k = 1, 2, \dots, K,$$

with

$$\hat{\bar{E}}_{k:\text{HT}} = \hat{\bar{Y}}_{k:\text{HT}} - \mathbf{b}_k' \hat{\mathbf{X}}_{\text{HT}} = \sum_{i \in s} \left(\frac{\mathbf{p}_{ik}' (\mathbf{y}_i - \mathbf{B}' \mathbf{x}_i)}{\pi_i N} \right),$$

and where \mathbf{B} is an $H \times K$ matrix of which the columns are the H -vectors \mathbf{b}_k . Now $\hat{\bar{\mathbf{Y}}}_{\text{GREG}} = (\hat{\bar{Y}}_{1:\text{greg}}, \dots, \hat{\bar{Y}}_{K:\text{greg}})'$ is proposed as an approximately design-unbiased estimator for $E_m \bar{\mathbf{Y}}$.

2.4 Variance Estimation of Treatment Effects

Let \mathbf{V} denote the covariance matrix of $\hat{\bar{\mathbf{Y}}}_{\text{GREG}}$. To estimate the covariance terms of \mathbf{V} , vectors \mathbf{y}_i containing the observations of all K treatments obtained from each sampling unit are required. Since in the experimental designs under consideration each sampling unit is assigned to one of the K treatments, only one of the components of \mathbf{y}_i , for $i \in s$, is actually observed. Consequently, a design-unbiased estimator for \mathbf{V} cannot be derived. Van den Brakel and Binder (2000, 2004) tried to overcome this problem by imputing the unobserved components. The usefulness of their results, however, depends on the correctness of the imputation model. In the present paper, this problem is circumvented by deriving a design-based estimator for \mathbf{CVC}' , i.e., the covariance matrix of the contrasts of $\hat{\bar{\mathbf{Y}}}_{\text{GREG}}$, which is sufficient for the Wald statistic (12).

Expressions for the generalized regression estimator are derived first. Results for the Horvitz-Thompson estimator are given as a special case. The covariance matrix of the contrasts of $\hat{\bar{\mathbf{Y}}}_{\text{GREG}}$ can be approximated by the covariance matrix of the contrasts of $\hat{\bar{\mathbf{E}}}_{\text{HT}} = (\hat{\bar{E}}_{1:\text{HT}}, \dots, \hat{\bar{E}}_{K:\text{HT}})'$. Let Cov_s and Cov_e denote the covariances with respect to the sample design and the experimental design respectively. Now, consider the following variance decomposition:

$$\begin{aligned} \mathbf{CVC}' &= \text{Cov}_m E_s E_e (\hat{\mathbf{C}}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) \\ &+ E_m \text{Cov}_s E_e (\hat{\mathbf{C}}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s) + E_m E_s \text{Cov}_e (\hat{\mathbf{C}}\hat{\bar{\mathbf{E}}}_{\text{HT}} | m, s). \end{aligned} \quad (21)$$

Since $E_e(\mathbf{p}_{ik}) = \mathbf{r}_k$ (see (42) in the appendix), it follows that

$$E_e(\hat{\mathbf{E}}_{\text{HT}}|m, s) = \sum_{i \in s} \left(\frac{(\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i)}{\pi_i N} \right). \quad (22)$$

Under the condition that a constant H -vector \mathbf{a} exists such that $\mathbf{a}'\mathbf{x}_i = 1$ for all $i \in U$, it is proven in the appendix that

$$\mathbf{C}(\mathbf{y}_i - \mathbf{B}'\mathbf{x}_i) = \mathbf{C}\mathbf{e}_i. \quad (23)$$

The stated condition implicitly assumes that the size of the finite population is known and is used as auxiliary information. This condition holds for weighting models that contain an intercept or one or more categorical variables that partition the population into subpopulations. Using model assumptions (2) and (3), it follows from (22) and (23) that

$$\begin{aligned} \text{Cov}_m E_s E_e(\hat{\mathbf{C}}\hat{\mathbf{E}}_{\text{HT}}|m, s) &= \text{Cov}_m \left(\frac{1}{N} \sum_{i=1}^N \mathbf{C}\mathbf{e}_i \right) \\ &= \frac{1}{N^2} \sum_{i=1}^N \mathbf{C}\Sigma_i \mathbf{C}', \end{aligned} \quad (24)$$

and

$$\begin{aligned} E_m \text{Cov}_s E_e(\hat{\mathbf{C}}\hat{\mathbf{E}}_{\text{HT}}|m, s) &= E_m \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N (\pi_{ii'} - \pi_i \pi_{i'}) \\ &\times \frac{\mathbf{C}\mathbf{e}_i \mathbf{e}_{i'}' \mathbf{C}'}{\pi_i \pi_{i'}} = \frac{1}{N^2} \sum_{i=1}^N \left(\frac{1}{\pi_i} - 1 \right) \mathbf{C}\Sigma_i \mathbf{C}'. \end{aligned} \quad (25)$$

For the third term in (21), it is proven in the appendix for an RBD that

$$\begin{aligned} E_m E_s \text{Cov}_e(\hat{\mathbf{C}}\hat{\mathbf{E}}_{\text{HT}}|m, s) &= E_m E_e(\mathbf{C}\mathbf{D}\mathbf{C}') \\ &- \frac{1}{N^2} \sum_{i=1}^N \frac{\mathbf{C}\Sigma_i \mathbf{C}'}{\pi_i}, \end{aligned} \quad (26)$$

where \mathbf{D} is a $K \times K$ diagonal matrix with diagonal elements

$$\begin{aligned} d_k &= \sum_{j=1}^J \frac{1}{n_{jk}} \frac{1}{n_{j+} - 1} \sum_{i \in s_j} \\ &\left(\frac{n_{j+}(\mathbf{y}_{ik} - \mathbf{b}_k'\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{n_{j+}(\mathbf{y}_{ik} - \mathbf{b}_k'\mathbf{x}_{i'})}{N\pi_{i'}} \right)^2 \equiv \sum_{j=1}^J \frac{S_{E_k}^2}{n_{jk}}. \end{aligned} \quad (27)$$

If the results obtained in (24), (25) and (26) are inserted in (21), then it follows that

$$\mathbf{CVC}' = E_m E_s \mathbf{C}\mathbf{D}\mathbf{C}'. \quad (28)$$

Conditionally on the realization of m and s , an approximately design-unbiased estimator for \mathbf{D} in (28) can be derived. Therefore, \mathbf{CVC}' can conveniently be stated implicitly as the expectation over the measurement error model and the sampling design. See Van den Brakel (2001) for explicit expressions for \mathbf{CVC}' . Given the realization of

m and s , the allocation of the sampling units within each block to the subsamples s_{jk} can be considered as simple random sampling without replacement from block s_j . Consequently, for an RBD it follows that an approximately design-unbiased estimator for \mathbf{D} is given by a $K \times K$ diagonal matrix $\hat{\mathbf{D}}$ with diagonal elements

$$\begin{aligned} \hat{d}_k &= \sum_{j=1}^J \frac{1}{n_{jk}} \frac{1}{n_{j+} - 1} \sum_{i=1}^{n_{j+}} \\ &\left(\frac{n_{j+}(\mathbf{y}_{ik} - \hat{\mathbf{b}}_k'\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{jk}} \sum_{i'=1}^{n_{jk}} \frac{n_{j+}(\mathbf{y}_{ik} - \hat{\mathbf{b}}_k'\mathbf{x}_{i'})}{N\pi_{i'}} \right)^2 \equiv \sum_{j=1}^J \frac{\hat{S}_{E_k}^2}{n_{jk}}. \end{aligned} \quad (29)$$

An approximately design-unbiased estimator for \mathbf{CVC}' in (28) is given by $\hat{\mathbf{C}}\hat{\mathbf{D}}\hat{\mathbf{C}}'$. Results for a CRD follow directly as a special case from (27) and (29) where $J=1$, $n_{j+} = n_+$ and $n_{jk} = n_k$. As an alternative, the residuals $(\mathbf{y}_{ik} - \hat{\mathbf{b}}_k'\mathbf{x}_i)$ in (29) can be multiplied by the correction weights (also called g -weights, Särndal *et al.* 1992, result 6.6.1). Since, \mathbf{CVC}' in (28) is defined implicitly as the expectation over the sampling design, (29) is approximately design-unbiased under general complex sampling schemes. This variance estimator only requires that the fraction of sampling units assigned to the different treatments according to the experimental design is fixed in advance. The size of the sample as well as the blocks might be random with respect to the sample design, *e.g.*, in the case of an RBD where clusters or PSU's are the block variable.

The variance estimator $\hat{\mathbf{C}}\hat{\mathbf{D}}\hat{\mathbf{C}}'$ has a structure as if the K subsamples had been drawn independently from each other, where the sampling units are selected with unequal probabilities (π_i/n_+) with replacement in the case of a CRD, or (π_i/n_{j+}) with replacement within each block j in the case of an RBD (compare (29) with Cochran 1977, equation (9A.16)). It is remarkable that the second order inclusion probabilities of the sampling design have vanished. This is caused by:

1. The assumption of additive treatment effects in the measurement error model, *i.e.*, β_k for all $i \in U$ observed under treatment k .
2. The assumption that measurement errors between individuals are independent.
2. A properly chosen weighting scheme such that the condition $\mathbf{a}'\mathbf{x}_i = 1$ for all $i \in U$ is satisfied.
4. The fact that variances are calculated for the contrasts between the subsample means.

The design variance of the first-order Taylor series approximation of the generalized regression estimator consists of the residuals $(\mathbf{y}_{ik} - \mathbf{b}_k'\mathbf{x}_i)$. From the proof of (23) it follows that under a weighting scheme that satisfies the condition

$\mathbf{a}'\mathbf{x}_i = 1$ for all $i \in U$, the treatment effects β_k vanish from the residuals $(y_{ik} - \mathbf{b}'_k \mathbf{x}_i)$ in (23). In these residuals three terms remain:

1. The residual of the linear regression model of the intrinsic value, i.e., $e_i = u_i - \mathbf{b}' \mathbf{x}_i$.
2. A term concerning the bias due to the interviewer effects. This term is equal to $\psi_{ii} - \mathbf{d}' \mathbf{x}_i$, where \mathbf{d} denotes the regression coefficients from the regression function of the interviewer effects on the auxiliary variables \mathbf{x}_i , (see proof of (23) in the appendix).
3. The measurement errors ε_{ik} .

The residuals of the intrinsic values e_i and the bias due to the interviewer effects do not depend on the different treatments and therefore cancel out in the contrasts of the residuals in (23). Only the measurement errors ε_i remain in the contrasts of the residuals in (23). As a result, the two terms $\text{Cov}_m \mathbf{E}_s \mathbf{E}_e (\mathbf{C}\hat{\mathbf{E}}_{\text{HT}} | m, s)$ and $\mathbf{E}_m \text{Cov}_s \mathbf{E}_e (\mathbf{C}\hat{\mathbf{E}}_{\text{HT}} | m, s)$ only contain the measurement errors ε_{ik} . Due to the assumption of independence of the measurement errors between individuals, the cross products between individuals, which contain the second order inclusion probabilities in (24) and (25) vanish. The covariance structure of the third term of (21) is mainly determined by the randomization mechanism of the experimental design. For a CRD this comes down to the selection of K subsamples from s by means of simple random sampling without replacement. For an RBD this comes down to the selection of K subsamples from s by means of stratified simple random sampling without replacement where strata correspond to the blocks of the experiment. In the variance of the contrasts of the subsample means, the finite population corrections in the design variance of the subsample means cancel out against the design covariance between the subsample means. As a result, the leading term of (26), i.e., $\mathbf{E}_m \mathbf{E}_s \mathbf{C} \mathbf{D} \mathbf{C}'$, has a structure as if the K subsamples were drawn independently of each other by means of simple random sampling with replacement in the case of a CRD, or stratified simple random sampling with replacement in the case of an RBD. Second order inclusion probabilities appear if the expectation with respect to the sampling design in (28) is made explicit, see Van den Brakel (2001).

The minimum use of auxiliary information is a weighting scheme where $\mathbf{x}_i = (1)$ and $\omega_i^2 = \omega^2$ for all $i \in U$. Under this weighting scheme it follows that

$$\hat{y}_{k:\text{greg}} = \left(\sum_{i=1}^{n_k} \frac{1}{\pi_i} \right)^{-1} \left(\sum_{i=1}^{n_k} \frac{y_{ik}}{\pi_i} \right) \equiv \tilde{y}_k, \quad (30)$$

which can be recognized as the ratio estimator for a population mean, originally proposed by Hájek (1971). It also follows that $\hat{\mathbf{b}}_k = (\tilde{y}_k)$ and that an approximately

design-unbiased estimator for the covariance matrix of the treatment effects is given by (29) with $\hat{\mathbf{b}}'_k \mathbf{x}_i = \tilde{y}_k$.

If $\sum_{i=1}^{n_k} 1/\pi_i \equiv \hat{N} = N$, then the ratio estimator (30) corresponds with the regular Horvitz-Thompson estimator. This condition is satisfied in the case of a CRD or an RBD embedded in a simple random sampling design, an RBD embedded in a stratified simple random sampling design where strata are used as block variables or a CRD embedded in a stratified simple random sampling design with proportional allocation. Under the condition $\hat{N} = N$, expressions for the design variance of the Horvitz-Thompson estimator are given by (27) and (29), where $y_{ik} - \mathbf{b}'_k \mathbf{x}_i$ and $y_{ik} - \hat{\mathbf{b}}'_k \mathbf{x}_i$ are replaced by y_{ik} . Variance expressions for the Horvitz-Thompson estimator are more complicated if $\hat{N} \neq N$, see Van den Brakel (2001).

2.5 The Wald Test

Inserting the design-unbiased estimators for the subsample means and the covariance matrix of the contrasts between these subsample means into (12) leads to the design-based Wald statistic

$$W = \hat{\mathbf{Y}}_{\text{GREG}}^t \mathbf{C}' (\mathbf{C} \hat{\mathbf{D}} \mathbf{C}')^{-1} \mathbf{C} \hat{\mathbf{Y}}_{\text{GREG}}. \quad (31)$$

It is proven in the appendix that this expression can be simplified to:

$$W = \sum_{k=1}^K \frac{\hat{y}_{k:\text{greg}}^2}{\hat{d}_k} - \frac{1}{\sum_{k=1}^K \frac{1}{\hat{d}_k}} \left(\sum_{k=1}^K \frac{\hat{y}_{k:\text{greg}}}{\hat{d}_k} \right)^2. \quad (32)$$

For general sampling schemes, the asymptotic distribution of this test statistic will be unknown. However, if the sampling design is simple random sampling without replacement and the experimental design is a CRD, then Lehmann (1975, appendix 8), based on the work of Hájek (1960), gives sufficient conditions under which $\hat{\mathbf{E}}_{\text{HT}}$ is asymptotically multivariate normal distributed with mean $\mathbf{E}_s \mathbf{E}_{\hat{k}} (\hat{\mathbf{E}}_{\text{HT}} | m, s) = \bar{\mathbf{E}}$ and covariance matrix $\hat{\mathbf{V}} = \text{Cov}_s \mathbf{E}_s (\hat{\mathbf{E}}_{\text{HT}} | m, s) + \mathbf{E}_s \text{Cov}_{\hat{k}} (\hat{\mathbf{E}}_{\text{HT}} | m, s)$ if $n_{+k} \rightarrow \infty$ and $(N - n_{+k}) \rightarrow \infty: (\hat{\mathbf{E}}_{\text{HT}} | m) \rightarrow N(\bar{\mathbf{E}}, \hat{\mathbf{V}})$. Hence, $(\mathbf{C}\hat{\mathbf{E}}_{\text{HT}} | m) \rightarrow N(\mathbf{C}\bar{\mathbf{E}}, \mathbf{C}\hat{\mathbf{V}}\mathbf{C}')$, with $\mathbf{C}\bar{\mathbf{E}} = (1/N) \sum_{i=1}^N \mathbf{C}\mathbf{e}_i$. Since the $\mathbf{C}\mathbf{e}_i$ are mutually independent random variables with means equal to zero and covariance matrix $\mathbf{C} \sum_i \mathbf{C}'$ we have by the ordinary central limit theorem $(\mathbf{C}\bar{\mathbf{E}}) \rightarrow N(0, (1/N^2) \sum_{i=1}^N \mathbf{C} \sum_i \mathbf{C}')$. Combining both limit distributions we obtain that unconditionally $\mathbf{C}\hat{\mathbf{E}}_{\text{HT}} \rightarrow N(0, \mathbf{C}\mathbf{V}\mathbf{C}')$ and thus $\mathbf{C}\hat{\mathbf{Y}}_{\text{GREG}} \rightarrow N(\mathbf{C}\bar{\boldsymbol{\beta}}, \mathbf{C}\mathbf{V}\mathbf{C}')$. As a result it follows under the null hypothesis that W is asymptotically chi-squared distributed with $K-1$ degrees of freedom (Searle 1971, theorem 2, chapter 2). For more complex sampling designs it is usually conjectured that

$\hat{\mathbf{C}}\hat{\mathbf{Y}}_{\text{GREG}} \rightarrow N(\mathbf{C}\boldsymbol{\beta}, \mathbf{CVC}')$. Then W is still asymptotically chi-squared distributed with $K - 1$ degrees of freedom. The validity of this conjecture has been confirmed by simulation studies, see section 4 and Van den Brakel (2001).

2.6 Pooled Variance Estimators

In the case of an RBD the n_{++} sampling units of s are divided into JK groups of size n_{jk} . For each of these JK subsamples separate population variances $\hat{S}_{E_{jk}}^2$ have to be estimated. If the number of experimental units n_{jk} available for the estimation of these population variances becomes too small, then these estimates might become unstable. In such situations, more stable estimates can be obtained by pooling estimates of the population variances within the blocks.

The residuals of the generalized regression estimator, $(y_{ik} - \mathbf{b}'_k \mathbf{x}_i)$, only depend on the k^{th} treatment effect through the measurement errors ε_{ik} . Under the assumption that $\Sigma_i = \sigma^2 \mathbf{I}$ in (3) for all $i \in U$, it follows that the $S_{E_{jk}}^2$ within each block are identical parameters, *i.e.*, $S_{E_{j1}}^2 = \dots = S_{E_{jK}}^2 = S_{E_j}^2$, for $j = 1, 2, \dots, J$. Under this assumption, it is efficient to use a pooled estimator for $S_{E_j}^2$;

$$\hat{S}_{E_j, P_1}^2 = \frac{1}{(n_{j+} - 1)} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \left(\frac{n_{j+}(y_{ik} - \mathbf{b}'_k \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{j+}} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \frac{n_{j+}(y_{ik} - \mathbf{b}'_k \mathbf{x}_i)}{N \pi_i} \right)^2 \quad (33)$$

or alternatively

$$\hat{S}_{E_j, P_2}^2 = \frac{1}{(n_{j+} - K)} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \left(\frac{n_{j+}(y_{ik} - \mathbf{b}'_k \mathbf{x}_i)}{N \pi_i} - \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} \frac{n_{j+}(y_{ik} - \mathbf{b}'_k \mathbf{x}_i)}{N \pi_i} \right)^2. \quad (34)$$

There are several special cases where the design-based Wald statistic coincides with the F -statistics known from more standard model-based analysis procedures. Consider an RBD embedded in a self-weighted sampling design where sampling units are allocated proportionally to the treatments over the blocks, *i.e.*, $\pi_i = n_{++}/N$ and $n_{jk}/n_{j+} = n_{+k}/n_{++}$ for all $j = 1, \dots, J$. Then, it follows from the results obtained for the ratio estimator (30) that $\hat{\mathbf{Y}}_{k, \text{greg}} = 1/n_{+k} \sum_{i=1}^{n_{+k}} y_{ik} \equiv \bar{y}_{+k}$ and $\mathbf{b}'_k \mathbf{x}_i = \bar{y}_{+k}$. Denote $\bar{y}_{j+} = 1/n_{j+} \sum_{i=1}^{n_{j+}} y_{ik}$ and $\bar{y}_{++} = 1/n_{++} \sum_{k=1}^K \sum_{i=1}^{n_{+k}} y_{ik}$, then it follows that

$$\begin{aligned} \frac{1}{n_{j+}} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} y_{ik} &= \bar{y}_{j+}, & \frac{1}{n_{j+}} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \mathbf{b}'_k \mathbf{x}_i &= \\ \frac{1}{n_{j+}} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \bar{y}_{+k} &= \sum_{k=1}^K \frac{n_{jk}}{n_{j+}} \bar{y}_{+k} = \sum_{k=1}^K \frac{n_{+k}}{n_{++}} \bar{y}_{+k} = \bar{y}_{++}. \end{aligned}$$

If $n_{j+} \approx n_{j+} - 1$, then it follows under the pooled variance estimator (33) that

$$\begin{aligned} \hat{d}_k &= \sum_{j=1}^J \frac{n_{j+}}{n_{jk}} \frac{n_{j+}}{n_{j+} - 1} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \left(\frac{y_{ik} - \mathbf{b}'_k \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \frac{y_{ik} - \mathbf{b}'_k \mathbf{x}_i}{N \pi_i} \right)^2 \\ &\approx \frac{1}{n_{+k}} \frac{1}{n_{++}} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ik} - \bar{y}_{+k} - \bar{y}_{j+} + \bar{y}_{++})^2 \equiv \frac{\hat{d}_{P_1}}{n_{+k}}. \quad (35) \end{aligned}$$

Denote $\bar{y}_{jk} = 1/n_{jk} \sum_{i=1}^{n_{jk}} y_{ik}$. Under the pooled variance estimator (34) it follows that

$$\begin{aligned} \hat{d}_k &= \sum_{j=1}^J \frac{n_{j+}}{n_{jk}} \frac{n_{j+}}{n_{j+} - K} \sum_{k=1}^K \sum_{i=1}^{n_{jk}} \left(\frac{y_{ik} - \mathbf{b}'_k \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} \frac{y_{ik} - \mathbf{b}'_k \mathbf{x}_i}{N \pi_i} \right)^2 \\ &\approx \frac{1}{n_{+k}} \frac{1}{n_{++}} \sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^{n_{jk}} (y_{ik} - \bar{y}_{jk})^2 \equiv \frac{\hat{d}_{P_2}}{n_{+k}}. \quad (36) \end{aligned}$$

Substituting these pooled variance estimators into the Wald statistic (32), leads to

$$W = \frac{1}{\hat{d}_{P_a}} \left(\sum_{k=1}^K n_{+k} (\bar{y}_{+k})^2 - n_{++} (\bar{y}_{++})^2 \right), \quad (37)$$

where \hat{d}_{P_a} is given by either (35) for $a=1$ or (36) for $a=2$. It can be recognized that $W/(K-1)$ in (37) with \hat{d}_{P_1} the pooled variance estimator (35), corresponds with the F -statistics of an ANOVA for a two-way layout without interactions. If \hat{d}_{P_2} (36) is inserted, then $W/(K-1)$ corresponds with the F -statistics of an ANOVA for a two-way layout with interactions (Scheffé 1959, chapter 4). A pooled variance estimator for a CRD follows as a special case from (35) and (36). Under both estimators it follows that $W/(K-1)$ corresponds with the F -statistics of the one-way ANOVA (Scheffé 1959, chapter 3).

2.7 Advantages of RBD's

The main advantage of RBD's is the elimination of the variation between the blocks in the analysis of treatment effects. Sampling units from the same stratum, PSU or cluster generally have a higher degree of homogeneity compared with sampling units from different strata, PSU's or clusters. This suggests using sampling structures like strata, PSU's or clusters as block variables in an RBD (Fienberg and Tanur 1987, 1988). Using these sampling structures as a block variable in an RBD, ensures that each stratum, PSU or cluster is sufficiently represented within

each subsample. Also interviewers are potential block variables, since this eliminates the variation in the observations due to fixed or random interviewer effects specified in measurement error model (1). For surveys where interviewers collect data by means of CAPI in separated geographical areas, blocking on interviewers also eliminates this regional variation from the target variable. The power of an experiment is maximized if sampling units are allocated proportionally to treatments over the blocks, *i.e.*, $n_{jk}/n_{j+} = n_{+k}/n_{++}$ for all $j = 1, \dots, J$ (see Van den Brakel 2001, chapter 6). This allocation is better preserved if interviewers are used as the block variables, since response rates between interviewers differ substantially. Unrestricted randomization by means of a CRD is not always feasible from a practical point of view. For example in CAPI surveys where interviewers collect data in geographical areas surrounding their places of residence, restricted randomization of sampling units within interviewers or geographical regions which are unions of adjacent interviewer regions might be required to avoid an unacceptable increase in the travel distance of the interviewers. This naturally leads to RBD's with interviewers or regions as block variables.

3. Design-Based Linear Regression Analysis

A design-based linear regression might be considered as an alternative for the analysis of embedded experiments. The observations are assumed to be the outcome of a linear regression model $y_i = B'x_i + e_i$, with x_i the vector containing Q explanatory variables, B the vector containing the regression coefficients, and e_i a residual. This model is mainly determined by the experimental design and contains the treatment factors, local control factors (*e.g.*, blocks) and covariates as explanatory variables (see *e.g.*, Montgomery 2001). Potential covariates are the auxiliary variables in the weighting scheme of the generalized regression estimator. The parameters of interest are the regression coefficients in the finite population, which are defined by $\beta = (X'X)^{-1}X'y$, where X is the $N \times Q$ design matrix of the experimental design, and y a N vector containing the observations obtained under the different treatments, as if the entire finite population is included in the experiment. The design matrix conceptually divides the population into K subpopulations or domains, which are observed under each of the K treatments of the experiment. The size of each subpopulation is determined by the fraction of sampling units assigned to each treatment in the experiment. A design-based estimator for the regression coefficients is given by $\hat{\beta} = (X'_n \Pi^{-1} X_n)^{-1} X'_n \Pi^{-1} y_n$, (Särndal *et al.* 1992, section 5.10). Here X_n is the $n \times Q$ design matrix, y_n a vector containing n observations obtained under the

different treatments of the n units included in the sample, and Π a $n \times n$ diagonal matrix containing the first order inclusion probabilities π_i of the sampling design. The approximate covariance matrix of $\hat{\beta}$, is given by (Särndal *et al.* 1992, section 5.10)

$$\text{Var}(\hat{\beta}) = (X'X)^{-1} \Lambda (X'X)^{-1}, \quad (38)$$

with $\Lambda = \text{Var}_\pi(X'_n \Pi^{-1} y_n - X'_n \Pi^{-1} X_n \beta)$. The elements of Λ are given by

$$\lambda_{qq'} = \sum_{i \in U} \sum_{i' \in U} (\pi_{ii'} - \pi_i \pi_{i'}) \frac{x_{iq} e_i}{\pi_i} \frac{x_{i'q'} e_{i'}}{\pi_{i'}}, \quad q, q' = 1, \dots, Q,$$

with $e_i = y_i - \beta'x_i$. Hypotheses about the subset of regression coefficients that reflect the treatment effects are tested with a Wald test, see *e.g.*, Skinner (1989).

The major drawback of this approach is that the estimation procedure doesn't account for the random assignment of sampling units to treatments according to the experimental design. In doing so the subsample estimates are erroneously treated as if they were domain estimates, which results in wrong design-variances. The covariance matrix of the treatment effects (28), derived in section 2.4, illustrates that the superimposition of the experimental design on the sampling design determines which specific features of the sampling design are nullified or preserved. For example, the effect of stratified sampling or two-stage sampling on the variance of the treatment effects is nullified under a CRD. This effect, however, is ignored by the linear regression approach, since $\text{Var}(\hat{\beta})$ only accounts for the variance of the sample design. Disregarding the experimental design in the variance estimation procedure becomes even more obvious under a complete enumeration of the finite population. Due to the experimental design, the entire finite population is randomly divided into K subsamples and the parameters under the different treatments are still estimated with a nonzero design variance. In this situation it follows for the linear regression approach that $\hat{\beta} = \beta$ and that $\text{Var}(\hat{\beta})$ is equal to zero because the design-variance induced by the experimental design is ignored. This contrasts with (28) that under a complete enumeration still reflects the design-variance due to the experimental design.

It is not immediately evident how the linear regression approach can be adjusted to allow for the randomization due to the sampling design as well as the experimental design. Conditionally on the realization of the sample, the experimental design can be described by first and second order inclusion probabilities. Let $\pi_{i|k}^k$ denote the first order inclusion probability that the i^{th} sampling unit is assigned to the k^{th} treatment and let $\pi_{i|k'}^{kk'}$ denote the second order inclusion probability that i^{th} sampling unit is assigned to the k^{th} treatment and the i'^{th} sampling unit is assigned to the k'^{th} treatment. A design-based estimator for β that accounts for the sampling design and the experimental

design is given by $\hat{\beta} = (\mathbf{X}_n' \mathbf{\Pi}^{*-1} \mathbf{X}_n)^{-1} \mathbf{X}_n' \mathbf{\Pi}^{*-1} \mathbf{y}_n$, where $\mathbf{\Pi}^*$ denotes the $n \times n$ diagonal matrix with first order inclusion probabilities $\pi_i^* = \pi_i \pi_{i|s}^*$. An approximation for the covariance matrix of $\hat{\beta}$ is given by (38), where Λ is obtained by conditioning on the realization of the sample, *i.e.*,

$$\Lambda = \text{Var}_s \text{E}_e (\mathbf{X}_n' \mathbf{\Pi}^{*-1} \mathbf{y}_n - \mathbf{X}_n' \mathbf{\Pi}^{*-1} \mathbf{X}_n \beta) \\ + \text{E}_s \text{Var}_e (\mathbf{X}_n' \mathbf{\Pi}^{*-1} \mathbf{y}_n - \mathbf{X}_n' \mathbf{\Pi}^{*-1} \mathbf{X}_n \beta).$$

This leads to the following expression for the elements of Λ :

$$\lambda_{qq'} = \sum_{i \in U} \sum_{i' \in U} (\pi_{ii'} - \pi_i \pi_{i'}) \frac{x_{iq} e_i}{\pi_i} \frac{x_{i'q'} e_{i'}}{\pi_{i'}} \\ + \sum_{i \in U} \sum_{i' \in U} \pi_{ii'} (\pi_{ii'|s}^{kk'} - \pi_{i|s}^k \pi_{i'|s}^{k'}) \frac{x_{iq} e_i}{\pi_i^*} \frac{x_{i'q'} e_{i'}}{\pi_{i'}^*},$$

which has the variance structure of a two-phase sample, where the first phase corresponds to the sampling design and the second phase to the experimental design. The sampling units are, according to the experimental design, assigned to only one of the K treatments. As a result it follows that $\pi_{ii'|s}^{kk'} = 0$ for $k \neq k'$, and $i = i'$, which hampers the derivation of an approximately design-unbiased estimator for the covariance terms of $\text{Var}(\hat{\beta})$, see also Van den Brakel and Binder (2000, 2004). In the analysis procedure proposed in section 2, this problem is circumvented by deriving a design-based estimator for the covariance matrix of the contrasts of $\hat{\mathbf{C}} \hat{\mathbf{Y}}_{\text{GREG}}$ instead of an estimator for the covariance matrix of $\hat{\mathbf{Y}}_{\text{GREG}}$ itself.

4. Simulation Study

In subsection 4.1, a simulation study is conducted to evaluate the performance of the design-based estimator for the covariance matrix of the contrasts between the subsample estimates $\hat{\mathbf{C}} \hat{\mathbf{D}} \hat{\mathbf{C}}'$ with diagonal elements (29) as well as the design-based Wald statistic W defined by (32) to test hypotheses about these contrasts. Subsequently, this design-based Wald test, the design-based linear regression approach and a standard ANOVA are applied to the analysis of a CRD and an RBD in subsection 4.2.

4.1 Evaluation of the Unbiasedness of $\hat{\mathbf{C}} \hat{\mathbf{D}} \hat{\mathbf{C}}'$ and the Distribution of W

In this simulation study, a measurement error model without interviewer effects is assumed, *i.e.*,

$$y_{ik} = u_i + \beta_k + \varepsilon_{ik}. \quad (39)$$

An artificial population consisting of 3 strata, 450 PSU's and 109,500 SSU's is generated by randomly drawing strictly positive values for the intrinsic values u_i of a target parameter. The sizes of the PSU's in the population are

unequal. The intrinsic values are generated in two steps. First, a positive value for each PSU in the population is drawn from a uniform distribution. Subsequently a positive value for each SSU, also drawn from a uniform distribution, is added to the value obtained for the PSU in the first step. Within each stratum different lower and upper boundaries and interval-widths for these uniform distributions are applied, such that the population can be stratified into three relatively homogeneous subpopulations. The intervals of the uniform distributions that are applied in the second step are smaller than the intervals of the uniform distributions in the first step. This resulted in a population where the intrinsic values for the SSU's within each PSU are clustered. The structure of the population is summarized in Table 1.

Table 1
Population

Stratum	Number of PSU's	Number of SSU's	Intrinsic value of target parameter			
			Mean	Std. dev.	Min. value	Max. value
1	70	6,250	22,183	12,001	7,607	50,915
2	130	18,250	6,128	1,866	3,007	10,490
3	250	85,000	1,407	732	512	3,248
Total	450	109,500	3,380	5,803	512	50,915

Samples are drawn repeatedly from this population by means of stratified two-stage sampling without replacement with unequal inclusion probabilities. The inclusion probabilities are chosen proportionally to the size of the target parameter. The sample sizes for the different strata are summarized in Table 2. For each sample, a new measurement error is generated for each population element. These measurement errors are drawn from a normal distribution with a mean equal to zero and a standard deviation proportional to the size of the intrinsic values. The range of the standard deviations varied from 1,000 for the SSU's with the largest intrinsic values in the first stratum to 10 for the SSU's with the smallest intrinsic values in the third stratum.

Table 2
Sample Design

Stratum	Number of PSU's	Number of SSU's
1	25	900
2	30	1,080
3	50	1,800
Total	105	3,780

Finally, the samples are randomly divided into four subsamples according to an experimental design, each with a size of 945 SSU's. Two different experimental designs are applied. In the first design, the SSU's are randomized over the four different treatments according to a CRD. In the second design, the SSU's are randomized over the four different treatments according to an RBD, where the three strata are used as the block variable. Within each block or stratum, 1/4 of the SSU's are randomly assigned to each treatment. Under both experimental designs, four different

sets of treatment effects are applied, one under the null hypothesis and three under different alternative hypotheses. This resulted in eight different simulations, which are specified in Table 3. Each simulation is based on $R=100,000$ resamples. Observations for the target parameter are obtained by adding a measurement error and a treatment effect to the intrinsic values according to (39).

Table 3
Summary of Simulation Settings

Experimental design		Treatment effects			
		β_1	β_2	β_3	β_4
CRD	RBD	0	0	0	0
CRD	RBD	0	20	40	60
CRD	RBD	0	40	80	120
CRD	RBD	0	80	160	240

The data obtained in each resample are analyzed with the extended Horvitz-Thompson estimator (30). Let \tilde{y}_k^r denote the subsample estimate obtained under the k^{th} treatment in the r^{th} resample. The vector with the four subsample estimates obtained in the r^{th} resample is denoted by $\tilde{\mathbf{Y}}^r = (\tilde{y}_1^r, \tilde{y}_2^r, \tilde{y}_3^r, \tilde{y}_4^r)'$. The vector with the three contrasts in the r^{th} resample is equal to $\mathbf{C}\tilde{\mathbf{Y}}^r$, with $\mathbf{C} = (\mathbf{j}'; -\mathbf{I})$, \mathbf{j} a vector of order 3 with each element equal to one, and \mathbf{I} the 3×3 identity matrix. Furthermore, \hat{d}_k^r denotes the diagonal elements of the estimated covariance matrix, obtained under the r^{th} resample. An expression for \hat{d}_k^r is given by (29) with $\hat{\mathbf{b}}_i' \mathbf{x}_i = \tilde{y}_i^r$. The estimated covariance matrix of the treatment effects is equal to $\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}'$, with $\hat{\mathbf{D}}^r = \text{diag}(\hat{d}_1^r, \hat{d}_2^r, \hat{d}_3^r, \hat{d}_4^r)$. Finally $W^r = (\mathbf{C}\tilde{\mathbf{Y}}^r)'(\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}')^{-1}(\mathbf{C}\tilde{\mathbf{Y}}^r)$ denotes the Wald statistic observed in the r^{th} resample. Based on the $R=100,000$ resamples within each simulation, the population parameters under the different treatments can be approximated by

$$\bar{\mathbf{Y}} = \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{Y}}^r,$$

with $\bar{\mathbf{Y}} = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4)'$. From (10) it follows that the real treatment effects in the measurement error model can be approximated by $\mathbf{C}\bar{\mathbf{Y}} \approx \mathbf{C}\boldsymbol{\beta}$. Furthermore, the mean of the estimated resample covariance matrices can be calculated as

$$\mathbf{C}\bar{\mathbf{D}}\mathbf{C}' = \frac{1}{R} \sum_{r=1}^R \mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}',$$

and the mean of the resample Wald statistics as

$$\bar{W} = \frac{1}{R} \sum_{r=1}^R W^r. \tag{40}$$

An approximation of the real covariance matrix of the treatment effects is given by

$$\mathbf{CVC}' = \frac{1}{R-1} \sum_{r=1}^R \mathbf{C}(\tilde{\mathbf{Y}}^r - \bar{\mathbf{Y}})(\tilde{\mathbf{Y}}^r - \bar{\mathbf{Y}})' \mathbf{C}'. \tag{41}$$

The performance of the variance estimation procedure is evaluated by comparing $\mathbf{C}\bar{\mathbf{D}}\mathbf{C}'$ to \mathbf{CVC}' . If the derived variance estimator $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}'$ is approximately design-unbiased, then the mean of resample covariance matrices $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}'$ must tend to the real covariance matrix \mathbf{CVC}' , for $R \rightarrow \infty$. An impression of the precision of the derived variance estimator is obtained by calculating the standard deviation of the elements of $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}'$, and is denoted by $\sigma(\mathbf{C}\hat{\mathbf{D}}\mathbf{C}')$. The diagonal elements of $\bar{\mathbf{D}}$ are denoted \bar{d}_k .

If $\mathbf{C}\tilde{\mathbf{Y}}_{\text{GREG}}^r \rightarrow N(\mathbf{C}\boldsymbol{\beta}, \mathbf{CVC}')$, then it follows that $W \rightarrow \chi^2_{[K-1]|\delta}$, with $K-1$ the number of degrees of freedom and $\delta = 1/2(\mathbf{C}\boldsymbol{\beta})'(\mathbf{CVC}')^{-1}(\mathbf{C}\boldsymbol{\beta})$ the non-centrality parameter of the chi-squared distribution. In the simulation study, the non-centrality parameter under the alternative hypotheses can be calculated by inserting (41) in the expression of δ . Subsequently, the power of the Wald statistic for a particular set of treatment effects can be calculated by $P(W) = P(\chi^2_{[K-1]|\delta} > \chi^2_{[1-\alpha][K-1]})$ where $\chi^2_{[1-\alpha][K-1]}$ denotes the $(1-\alpha)^{\text{th}}$ percentile point of the central chi-squared distribution with $K-1$ degrees of freedom. The performance of the Wald statistic is evaluated by comparing $P(W)$ with the simulated power, which is defined as the fraction of significant runs observed in the R resamples, i.e.,

$$P^{\text{sim}}(W) = \frac{1}{R} \sum_{r=1}^R I(W^r > \chi^2_{[1-\alpha][K-1]}),$$

where $I(B)$ denotes the indicator variable which is equal to one if B is true, and equal to zero otherwise. The results of the simulations are summarized in Tables 4.1 through 4.8.

The means of the subsample estimates \bar{Y}_k under the null hypotheses in Tables 4.1 and 4.5 slightly overestimate the population mean in Table 1. This difference can be attributed to the bias of the extended Horvitz-Thompson estimator. The means of the contrasts between the subsample estimates $\mathbf{C}\bar{\mathbf{Y}}$, however, almost perfectly agree with the real treatment effects $\mathbf{C}\boldsymbol{\beta}$. The means of the resample covariance matrices $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}'$ tend to the values of the real covariance matrices \mathbf{CVC}' , which illustrates that the variance estimation procedure, derived in section 2.4, is approximately design-unbiased. The relative precision of the diagonal elements of $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}'$ is about 10.5% under this particular sample size. The simulated power based on the resample distribution of the Wald statistic approximates the real power reasonably well. On the average the simulated power is slightly higher. The expected value of the chi-squared distribution is equal to $E(\chi^2_{[K-1]|\delta}) = (K-1) + 2\delta$ (Searle 1971, section 2.4.h). If the resample distribution of the Wald statistic tends to a $\chi^2_{[K-1]|\delta}$, then the mean of the resample Wald statistics \bar{W} (40) must tend to the expected value of the chi-squared distribution. Indeed, it follows from Tables 4.1–4.8 that $\bar{W} \approx (K-1) + 2\delta$. Moreover, the

hypothesis that the resample distribution of the Wald statistic under the null hypothesis is equal to the central chi-squared distribution, is tested with the one-sample Kolmogorov-Smirnov test. This hypothesis is not rejected at a significance level of 5% for either the CRD or the RBD, and confirms the conjecture that the Wald statistic is asymptotically chi-squared distributed under stratified two-

stage sampling without replacement, unequal inclusion probabilities, and relatively large sampling fractions. If the simulations under a CRD are compared to an RBD, then it follows that blocking on strata results in a substantial increase of the precision of the estimated contrasts and the power of the tests in this particular situation.

Table 4.1
Simulation Results CRD, $\beta = (0, 0, 0, 0)^t$

Subsamples				Contrasts					Wald statistic		
k	β_k	\bar{Y}_k	\bar{d}_k	$k - k'$	$C\bar{Y}$	Diagonal elements of			α	$P(W)$	$P^{\text{sim}}(W)$
						CVC^t	$C\bar{D}C^t$	$\sigma(C\bar{D}C^t)$			
1	0	3,392	14,311						0.050	0.05000	0.05072
2	0	3,392	14,305	1 - 2	0	28,725	28,616	3,019	0.025	0.02500	0.02506
3	0	3,392	14,306	1 - 3	0	28,892	28,616	3,019	0.010	0.01000	0.01017
4	0	3,390	14,292	1 - 4	2	28,787	28,603	3,019	$\bar{W} : 3.01591$		$\delta : 0.0000$

Table 4.2
Simulation Results CRD, $\beta = (0, 20, 40, 60)^t$

Subsamples				Contrasts					Wald statistic		
k	β_k	\bar{Y}_k	\bar{d}_k			Diagonal elements of			α	$P(W)$	$P^{\text{sim}}(W)$
				$k - k'$	$C\bar{Y}$	CVC^t	$C\bar{D}C^t$	$\sigma(C\bar{D}C^t)$			
1	0	3,392	14,307						0.050	0.05842	0.05925
2	20	3,412	14,307	1 - 2	-20	28,635	28,614	3,026	0.025	0.03008	0.03040
3	40	3,432	14,314	1 - 3	-40	28,918	28,620	3,033	0.010	0.01257	0.01255
4	60	3,450	14,291	1 - 4	-58	28,624	28,597	3,025	$\bar{W} : 3.14037$		$\delta : 0.0697$

Table 4.3
Simulation Results CRD, $\beta = (0, 40, 80, 120)^t$

Subsamples				Contrasts					Wald statistic		
k	β_k	\bar{Y}_k	\bar{d}_k			Diagonal elements of			α	$P(W)$	$P^{\text{sim}}(W)$
				$k - k'$	$C\bar{Y}$	CVC^t	$C\bar{D}C^t$	$\sigma(C\bar{D}C^t)$			
1	0	3,392	14,314						0.050	0.08503	0.08523
2	40	3,432	14,307	1 - 2	-40	28,597	28,621	3,020	0.025	0.04704	0.04760
3	80	3,472	14,307	1 - 3	-80	28,947	28,622	3,022	0.010	0.02150	0.02165
4	120	3,511	14,295	1 - 4	-119	28,713	28,609	3,021	$\bar{W} : 3.55406$		$\delta : 0.2783$

Table 4.4
Simulation Results CRD, $\beta = (0, 80, 160, 240)^t$

Subsamples				Contrasts					Wald statistic		
k	β_k	\bar{Y}_k	\bar{d}_k			Diagonal elements of			α	$P(W)$	$P^{\text{sim}}(W)$
1	0	3,392	14,306	$k - k'$	$C\bar{Y}$	CVC^t	$C\bar{D}C^t$	$\sigma(C\bar{D}C^t)$	0.050	0.21198	0.2116
2	80	3,472	14,310	1 - 2	-80	28,748	28,616	3,026	0.025	0.13809	0.13885
3	160	3,552	14,312	1 - 3	-160	28,784	28,618	3,030	0.010	0.07703	0.07781
4	240	3,631	14,291	1 - 4	-239	28,538	28,598	3,022	$\bar{W} : 5.22065$		$\delta : 1.1203$

Table 4.5
Simulation Results RBD, $\beta = (0, 0, 0, 0)^t$

Subsamples				Contrasts					Wald statistic		
k	β_k	\bar{Y}_k	\bar{d}_k			Diagonal elements of			α	$P(W)$	$P^{\text{sim}}(W)$
				$k - k'$	$C\bar{Y}$	CVC^t	$C\bar{D}C^t$	$\sigma(C\bar{D}C^t)$			
1	0	3,389	3,088						0.050	0.05000	0.05168
2	0	3,389	3,088	1 - 2	0	6,175	6,176	647	0.025	0.02500	0.02640
3	0	3,389	3,088	1 - 3	0	6,216	6,176	647	0.010	0.01000	0.01060
4	0	3,389	3,088	1 - 4	0	6,217	6,176	647	$\bar{W} : 3.01483$		$\delta : 0.0000$

4.2 Comparison of Three Analysis Procedures

Furthermore, three possible analysis procedures for embedded experiments are compared, *i.e.*, the design-based Wald test proposed in section 2, a standard ANOVA where all observations are equally weighted and assumed to be i.i.d., and the design-based linear regression approach described in section 3. To this end two samples, each with a size of 3,780 SSU's are drawn from the finite population specified in Table 1, by means of the stratified two-stage sample design, which was also used in the previous simulation (see Table 2). For one sample, the SSU's are randomly divided into four subsamples, each with a size of 945, by means of a CRD. For the other sample the SSU's are randomly divided into four subsamples, each with a size of 945, by means of an RBD where the strata are used as the block variables. Both experiments are conducted under the alternative hypothesis where the treatment effects in the finite population are equal to $\beta = (0, 80, 160, 240)'$. The design-based linear regression analysis is performed with

Stata's SVYREG procedure that accounts for the stratification, two-stage sampling and the unequal selection probabilities of the sampling design (StataCorp. 2001). The ANOVA is performed with Stata's ANOVA procedure (StataCorp. 2001). The analysis results under a CRD are summarized for the design-based Wald test in Table 5.1, for the design-based linear regression approach in Table 5.2, and for the ANOVA in Table 5.3. Similarly, the analysis results under an RBD are summarized in Tables 6.1, 6.2, and 6.3.

As emphasized in section 3, the linear regression approach ignores the design variance due to the randomization of the sampling units over the subsamples with respect to the experimental design. As a result the standard errors of the treatment effects are smaller under the linear regression approach than in the case of the design-based Wald test, and the design-based regression approach results in smaller p-values for the test of treatment effects.

Table 4.6
Simulation Results RBD, $\beta = (0, 20, 40, 60)'$

Subsamples				Contrasts			Wald statistic		
<i>k</i>	β_k	\bar{Y}_k	\bar{d}_k	Diagonal elements of			α	$P(W)$	$P^{sim}(W)$
				$k - k'$	$C\bar{Y}$	CVC^t $C\bar{D}C^t$ $\sigma(C\bar{D}C^t)$			
1	0	3,390	3,090	$1 - 2$	-20	6,225 6,180 648	0.050	0.09099	0.09371
2	20	3,410	3,089	$1 - 3$	-40	6,177 6,181 648	0.025	0.05096	0.05238
3	40	3,430	3,090	$1 - 4$	-60	6,184 6,180 649	0.010	0.02365	0.02405
4	60	3,450	3,090				$\bar{W} : 3.66771$		$\delta : 0.3226$

Table 4.7
Simulation Results RBD, $\beta = (0, 40, 80, 120)'$

Subsamples				Contrasts			Wald statistic		
<i>k</i>	β_k	\bar{Y}_k	\bar{d}_k	Diagonal elements of			α	$P(W)$	$P^{sim}(W)$
				$k - k'$	$C\bar{Y}$	CVC^t $C\bar{D}C^t$ $\sigma(C\bar{D}C^t)$			
1	0	3,389	3,088	$1 - 2$	-40	6,178 6,176 647	0.050	0.23999	0.24310
2	40	3,429	3,088	$1 - 3$	-80	6,183 6,176 649	0.025	0.15999	0.16302
3	80	3,469	3,088	$1 - 4$	-120	6,189 6,176 649	0.010	0.09181	0.09458
4	120	3,509	3,088				$\bar{W} : 5.62182$		$\delta : 1.2905$

Table 4.8
Simulation Results RBD, $\beta = (0, 80, 160, 240)'$

Subsamples				Contrasts			Wald statistic		
<i>k</i>	β_k	\bar{Y}_k	\bar{d}_k	Diagonal elements of			α	$P(W)$	$P^{sim}(W)$
				$k - k'$	$C\bar{Y}$	CVC^t $C\bar{D}C^t$ $\sigma(C\bar{D}C^t)$			
1	0	3,390	3,091	$1 - 2$	-80	6,204 6,180 648	0.050	0.77340	0.77712
2	80	3,470	3,090	$1 - 3$	-160	6,210 6,181 648	0.025	0.68135	0.68789
3	160	3,550	3,090	$1 - 4$	-240	6,214 6,181 648	0.010	0.55796	0.56701
4	240	3,630	3,090				$\bar{W} : 13.48594$		$\delta : 5.1331$

Table 5.1
Design-based Wald Statistic, CRD

Subsamples			Contrasts			Wald statistic		
<i>k</i>	β_k	\tilde{y}_k	$k - k'$	$\tilde{y}_k - \tilde{y}_{k'}$	$\sqrt{\hat{\Delta}_k + \hat{\Delta}_{k'}}$	<i>W</i>	<i>df</i>	p-value
1	0	3,414	$1 - 2$	-124	164.915	2.4740	3	0.480
2	80	3,538	$1 - 3$	-182	162.542			
3	160	3,596	$1 - 4$	-249	164.782			
4	240	3,663						

Table 5.2
Design-based Regression, CRD

Source	Coefficient	Std. err.	Wald statistic	df	p-value
treatment			2.907	3	0.4062
treatment 1	- 182.14	177.60			
treatment 2	- 58.36	175.56			
treatment 4	66.79	170.46			
constant	3,596.47	194.75			

Table 5.3
Standard ANOVA, CRD

<i>k</i>	β_k	\bar{y}_k	Contrast		ANOVA				
			$k - k'$	$\bar{y}_k - \bar{y}_{k'}$	Source	df	MS	F	p-value
1	0	8021	1 - 2	- 73	Between treatments	3	14,432,816	0.14	0.9376
2	80	8094	1 - 3	66	Residual	3,776	104,924,668		
3	160	7955	1 - 4	- 221	Total	3,779			

Table 6.1
Design-based Wald Statistic, RBD

Subsamples			Contrasts		Wald statistic			
<i>k</i>	β_k	\tilde{y}_k	$k - k'$	$\tilde{y}_k - \tilde{y}_{k'}$	$\sqrt{\hat{d}_k + \hat{d}_{k'}}$	W	df	p-value
1	0	3,395	1 - 2	- 25	81.247	9.93011	3	0.0192
2	80	3,420	1 - 3	- 120	80.697			
3	160	3,515	1 - 4	- 231	82.383			
4	240	3,626						

Table 6.2
Design-based Regression, RBD

Source	Coefficient	Std.err.	Wald statistic	df	p-value
Block					
Block 2	-17,068.28	2,556.46			
Block 3	-21,999.39	2,540.98			
Treatment			18.4212	3	0.00036
Treatment 1	-211.51	74.84			
Treatment 2	-246.78	60.05			
Treatment 3	-97.91	73.39			
Constant	23,589.64	2543.25			

Table 6.3
Standard ANOVA, RBD

<i>k</i>	β_k	\bar{y}_{+k}	Contrast		ANOVA				
			$k - k'$	$\bar{y}_{+k} - \bar{y}_{+k'}$	Source	df	MS	F	p-value
1	0	8,815	1 - 2	665	Between blocks	2	1.6773 E+11		
2	80	8,150	1 - 3	249	Between treatments	3	84,377,227	1.99	0.1126
3	160	8,566	1 - 4	69	Residual	3,774	42,310,035		
4	240	8,746			Total	3,779	131,089,505		

The standard ANOVA is a naive approach, since it ignores the stratification, clustering and selection of sampling units using inclusion probabilities that are chosen proportional to the value of the target parameter. The net result of ignoring these aspects of the sampling design in the analysis is a severe over-estimation of the subsample estimates as well as the standard errors. Compared to the other two design-based procedures, this results in larger p-values for the test of treatment effects.

Another important advantage of the design-based Wald test compared to the design-based linear regression approach is that the Wald test always concerns the differences between the subsample estimates, which facilitate the interpretation of the results. This property is particularly important for embedded experiments aimed at the quantification of trend disruptions in the parameters of a survey due to adjustments in the survey design. In the case of a CRD, the linear regression model consists of one intercept parameter and three coefficients for the treatment effects. In

this particularly simple situation, the coefficients for the treatment effects are exactly equal to the differences between the subsample estimates. This property, however, doesn't hold for the treatment effects obtained under more complicated models, as for example in the case of the RBD.

5. Discussion and Conclusions

In this paper we discuss how the statistical methodology of randomized experiments and random survey sampling can support the design and analysis of experiments embedded in ongoing sample surveys. The sample survey design forms a prior framework for the application of principles, known from the theory of experimental designs, like randomization and local control by means of blocking on strata, PSU's, clusters or interviewers. To test hypotheses about the estimates of finite population parameters observed under different treatments of the experiment, a design-based Wald statistic for the analysis of CRD's and RBD's embedded in general complex sampling designs is derived using the Horvitz-Thompson estimator and the generalized regression estimator. The application of randomized sampling from a finite population in combination with this design-based analysis procedure enables us to generalize the results of the experiment observed in the specific sample to the entire survey population.

Since we allow for general complex sampling designs, a rather complicated expression for the covariance matrix of the treatment effects with nonzero off-diagonal entries is expected. The derived estimator for this covariance matrix, however, has a structure as if the sampling units were drawn with replacement and with unequal selection probabilities. No second order inclusion probabilities or design-covariances between the treatment effects are required, which simplifies the analysis considerably. For example, in the case of simple random sampling without replacement this result entails that the finite population correction factor should be disregarded in estimating the variance of contrasts. As a result a Wald statistic, derived from a design-based perspective under general complex sampling designs, is obtained that still has the appealing relatively simple structure of standard model-based analysis procedures.

For CRD's and RBD's embedded in a self-weighted sampling design analyzed with the extended Horvitz-Thompson estimator and a pooled variance estimator, the Wald statistic coincides with the F -statistic of an ANOVA for the one-way and two-way layouts. For the analysis of the embedded two-treatment experiment, a design-based version of the t -statistic can be derived as a special case of the Wald statistic. Expressions and more details about this design-based t -statistic and its relationship with Welch's t -statistic and the standard

t -statistic can be found in Van den Brakel and Renssen (1998), Van den Brakel (2001) or Van den Brakel and Van Berkel (2002).

The analysis procedure proposed in this paper is implemented in a software package, called X-tool. This tool will become available as a component of the Blaise survey processing software package, developed by Statistics Netherlands.

Appendix

Properties of the randomization vectors \mathbf{p}_{ik}

For CRD's and RBD's the randomization vectors \mathbf{p}_{ik} are defined by (14) and (15). As a consequence of the randomization mechanism of the experimental design, the vectors \mathbf{p}_{ik} are random with the following conditional probability mass functions. For a CRD we have

$$P\left(\mathbf{p}_{ik} = \frac{n_+}{n_k} \mathbf{r}_k \mid s\right) = \frac{n_k}{n_+} \quad \text{and} \quad P(\mathbf{p}_{ik} = \mathbf{0} \mid s) = 1 - \frac{n_k}{n_+}.$$

For an RBD we have

$$P\left(\mathbf{p}_{ik} = \frac{n_{j+}}{n_{jk}} \mathbf{r}_k \mid s_j\right) = \frac{n_{jk}}{n_{j+}} \quad \text{and} \quad P(\mathbf{p}_{ik} = \mathbf{0} \mid s_j) = 1 - \frac{n_{jk}}{n_{j+}}.$$

Properties of these vectors are derived for an RBD. Properties for a CRD follow as a special, since a CRD can be considered as an RBD with one block. Let w. pr. denote "with probability".

$$\mathbf{p}_{ik} \mathbf{p}_{ik}^t = \begin{cases} \left(\frac{n_{j+}}{n_{jk}}\right)^2 \mathbf{r}_k \mathbf{r}_k^t & \text{w. pr.: } \frac{n_{jk}}{n_{j+}} \quad \text{if } i \in s_j \\ \mathbf{O} & \text{w. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \end{cases}$$

$$\mathbf{p}_{ik} \mathbf{p}_{ik'}^t = \begin{cases} \frac{n_{j+}}{n_{jk}} \frac{n_{j'+}}{n_{jk'}} \mathbf{r}_k \mathbf{r}_{k'}^t & \text{w. pr.: } 0 \quad \text{if } i \in s_j \\ \mathbf{O} & \text{w. pr.: } 1 \end{cases}$$

$$\mathbf{p}_{ik} \mathbf{p}_{i'k'}^t = \begin{cases} \frac{n_{j+}}{n_{jk}} \frac{n_{j'+}}{n_{jk'}} \mathbf{r}_k \mathbf{r}_{k'}^t & \text{w. pr.: } \frac{n_{jk}}{n_{j+}} \frac{n_{jk'}}{(n_{j+} - 1)}, \text{ if } i \in s_j, i' \in s_{j'} \\ \mathbf{O} & \text{w. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{jk'}}{(n_{j+} - 1)}, \text{ if } i \in s_j, i' \in s_{j'} \\ \frac{n_{j+}}{n_{jk}} \frac{n_{j'+}}{n_{jk'}} \mathbf{r}_k \mathbf{r}_{k'}^t & \text{w. pr.: } \frac{n_{jk}}{n_{j+}} \frac{n_{jk'}}{n_{j'+}}, \text{ if } i \in s_j, i' \in s_{j'} \\ \mathbf{O} & \text{w. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{jk'}}{n_{j'+}}, \text{ if } i \in s_j, i' \in s_{j'} \end{cases}$$

$$\mathbf{p}_{ik} \mathbf{p}_{i'k}' = \begin{cases} \left(\frac{n_{j+}}{n_{jk}} \right)^2 \mathbf{r}_k \mathbf{r}_k' & \text{w. pr.: } \frac{n_{jk}}{n_{j+}} \frac{(n_{jk}-1)}{(n_{j+}-1)}, \text{ if } i \in s_j, i' \in s_j \\ \mathbf{O} & \text{w. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{(n_{jk}-1)}{(n_{j+}-1)}, \text{ if } i \in s_j, i' \in s_{j'} \\ \frac{n_{j+}}{n_{jk}} \frac{n_{j'+}}{n_{j'k}} \mathbf{r}_k \mathbf{r}_k' & \text{w. pr.: } \frac{n_{jk}}{n_{j+}} \frac{n_{j'k}}{n_{j'+}}, \text{ if } i \in s_j, i' \in s_{j'} \\ \mathbf{O} & \text{w. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{j'k}}{n_{j'+}}, \text{ if } i \in s_j, i' \in s_{j'} \end{cases}$$

The expectation of \mathbf{p}_{ik} with respect to the experimental design is given by:

$$E_e(\mathbf{p}_{ik}) = P\left(\mathbf{p}_{ik} = \frac{n_{j+}}{n_{jk}} \mathbf{r}_k\right) \frac{n_{j+}}{n_{jk}} \mathbf{r}_k + P(\mathbf{p}_{ik} = \mathbf{0}) \mathbf{0} = \mathbf{r}_k. \quad (42)$$

The following covariances with respect to the experimental design can be derived:

$$\text{Cov}_e(\mathbf{p}_{ik} \mathbf{p}_{i'k}') = \frac{(n_{j+} - n_{jk})}{n_{jk}} \mathbf{r}_k \mathbf{r}_k' \quad (43)$$

$$\text{Cov}_e(\mathbf{p}_{ik} \mathbf{p}_{i'k}') = -\mathbf{r}_k \mathbf{r}_k' \quad (44)$$

$$\text{Cov}_e(\mathbf{p}_{ik} \mathbf{p}_{i'k}') = \begin{cases} \frac{1}{(n_{j+}-1)} \mathbf{r}_k \mathbf{r}_k' & \text{if } i \in s_j \text{ and } i' \in s_j \\ \mathbf{O} & \text{if } i \in s_j \text{ and } i' \in s_{j'} \end{cases} \quad (45)$$

$$\begin{aligned} & \text{Cov}_e(\mathbf{p}_{ik} \mathbf{p}_{i'k}') \\ &= \begin{cases} -\frac{(n_{j+} - n_{jk})}{n_{jk}} \frac{1}{(n_{j+} - 1)} \mathbf{r}_k \mathbf{r}_k' & \text{if } i \in s_j \text{ and } i' \in s_j \\ \mathbf{O} & \text{if } i \in s_j \text{ and } i' \in s_{j'} \end{cases} \end{aligned} \quad (46)$$

Proof of formula (23)

Under the stated condition that a constant H -vector \mathbf{a} exists such that $\mathbf{a}' \mathbf{x}_i = 1$ for all $i \in U$, and conditional on the realization of $u_i, i=1, \dots, N$, according to superpopulation model (16), it follows that $\tilde{\mathbf{b}}_k$ in (18) can be evaluated as

$$\begin{aligned} E_m(\tilde{\mathbf{b}}_k) &= E_m \left(\sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i'}{\omega_i^2} \right)^{-1} \sum_{i=1}^N \frac{\mathbf{x}_i y_i}{\omega_i^2} \\ &= \left(\sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i'}{\omega_i^2} \right)^{-1} \sum_{i=1}^N \frac{\mathbf{x}_i (u_i + \Psi_i)}{\omega_i^2} + \mathbf{a} \beta_k \\ &= \mathbf{b} + \mathbf{d} + \mathbf{a} \beta_k, \end{aligned} \quad (47)$$

where \mathbf{b} denotes the regression coefficients defined by (17) and \mathbf{d} denotes the regression coefficients from the regression function of the interviewer effects on the auxiliary variables \mathbf{x}_i . From result (47) it follows that

$\mathbf{B}' \mathbf{x}_i = \mathbf{j}(\mathbf{b}' \mathbf{x}_i + \mathbf{d}' \mathbf{x}_i) + \beta$. Since $\mathbf{C} \mathbf{j} = \mathbf{0}$ and from measurement error model (1) and linear regression model (16) it follows that

$$\begin{aligned} \mathbf{C}(\mathbf{y}_i - \mathbf{B}' \mathbf{x}_i) &= \mathbf{C}(\mathbf{j} u_i + \mathbf{j} \Psi_i + \beta + \varepsilon_i - \mathbf{j}(\mathbf{b}' + \mathbf{d}') \mathbf{x}_i - \beta) \\ &= \mathbf{C} \varepsilon_i, \quad \text{QED.} \end{aligned}$$

Proof of formula (26) for an RBD

First an expression for $\text{Cov}_e(\hat{\mathbf{C}}_{\text{HT}} | m, s)$ is derived. Let $\mathbf{e}_i = (e_{i1}, \dots, e_{iK})'$ denote a K -vector with elements $e_{ik} = y_{ik} - \mathbf{b}_k' \mathbf{x}_i$. Consequently, $\mathbf{e}_i = \mathbf{y}_i - \mathbf{B}' \mathbf{x}_i$. Note that $E_m E_s \text{Cov}_e(\hat{\mathbf{C}}_{\text{HT}} | m, s) = \mathbf{C} E_m E_s \text{Cov}_e(\hat{\mathbf{E}}_{\text{HT}} | m, s) \mathbf{C}'$ with $\hat{\mathbf{E}}_{\text{HT}} = (\hat{E}_{1;\text{HT}}, \dots, \hat{E}_{K;\text{HT}})'$. Furthermore note that

$$\hat{E}_{k;\text{HT}} = \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{p}_{ik} (\mathbf{y}_i - \mathbf{B}' \mathbf{x}_i)}{\pi_i N} \right) = \sum_{i=1}^{n_{j+}} \frac{\mathbf{p}_{ik} \mathbf{e}_i}{\pi_i N}. \quad (48)$$

Using (43) and (46), the diagonal elements of $\text{Cov}_e(\hat{\mathbf{E}}_{\text{HT}} | m, s)$ can be elaborated as

$$\begin{aligned} & \text{Var}_e(\hat{E}_{k;\text{HT}} | m, s) \\ &= \text{Cov}_e \left(\sum_{i=1}^{n_{j+}} \frac{\mathbf{p}_{ik} \mathbf{e}_i}{\pi_i N}, \sum_{i'=1}^{n_{j+}} \frac{\mathbf{p}_{i'k} \mathbf{e}_{i'}}{\pi_{i'} N} \middle| m, s \right) \\ &= \sum_{j=1}^J \left(\sum_{i=1}^{n_{j+}} \frac{\mathbf{e}_i'}{\pi_i N} \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k}' | m, s) \frac{\mathbf{e}_i}{\pi_i N} \right. \\ & \quad \left. + \sum_{i=1}^{n_{j+}} \sum_{i' \neq i=1}^{n_{j+}} \frac{\mathbf{e}_i'}{\pi_i N} \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k}' | m, s) \frac{\mathbf{e}_{i'}}{\pi_{i'} N} \right) \\ &= \sum_{j=1}^J \left(\frac{n_{j+}}{(n_{j+}-1)} \frac{n_{j+}}{n_{jk}} \sum_{i=1}^{n_{j+}} \left(\frac{e_{ik}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik}}{\pi_i N} \right)^2 \right. \\ & \quad \left. - \frac{n_{j+}}{(n_{j+}-1)} \sum_{i=1}^{n_{j+}} \left(\frac{e_{ik}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik}}{\pi_i N} \right)^2 \right) \end{aligned} \quad (49)$$

Using (44) and (45), the off-diagonal elements of $\text{Cov}_e(\hat{\mathbf{E}}_{\text{HT}} | m, s)$ can be elaborated as

$$\begin{aligned} & \text{Cov}_e(\hat{E}_{k;\text{HT}}, \hat{E}_{k';\text{HT}} | m, s) \\ &= \text{Cov}_e \left(\sum_{i=1}^{n_{j+}} \frac{\mathbf{p}_{ik} \mathbf{e}_i}{\pi_i N}, \sum_{i'=1}^{n_{j+}} \frac{\mathbf{p}_{i'k'} \mathbf{e}_{i'}}{\pi_{i'} N} \middle| m, s \right) \\ &= \sum_{j=1}^J \left(\sum_{i=1}^{n_{j+}} \frac{\mathbf{e}_i'}{\pi_i N} \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k'}' | m, s) \frac{\mathbf{e}_i}{\pi_i N} \right. \\ & \quad \left. + \sum_{i=1}^{n_{j+}} \sum_{i' \neq i=1}^{n_{j+}} \frac{\mathbf{e}_i'}{\pi_i N} \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k'}' | m, s) \frac{\mathbf{e}_{i'}}{\pi_{i'} N} \right) \\ &= \sum_{j=1}^J - \frac{n_{j+}}{(n_{j+}-1)} \sum_{i=1}^{n_{j+}} \left(\frac{e_{ik}}{\pi_i N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{ik}}{\pi_i N} \right) \\ & \quad \left(\frac{e_{i'k'}}{\pi_{i'} N} - \frac{1}{n_{j+}} \sum_{i=1}^{n_{j+}} \frac{e_{i'k'}}{\pi_{i'} N} \right). \end{aligned} \quad (50)$$

The results (49) and (50) can be written in matrix notation;

$$\begin{aligned} \text{Cov}_e(\hat{\mathbf{E}}_{\text{HT}} | m, s) \\ = \mathbf{D} - \sum_{j=1}^J \frac{n_{j+}}{n_{j+} - 1} \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{y}_{ik} - \mathbf{B}'_k \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{y}_{i'k} - \mathbf{B}'_k \mathbf{x}_{i'}}{N \pi_{i'}} \right) \\ \left(\frac{\mathbf{y}_{ik} - \mathbf{B}'_k \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{y}_{i'k} - \mathbf{B}'_k \mathbf{x}_{i'}}{N \pi_{i'}} \right)' \end{aligned}$$

where \mathbf{D} denotes a $K \times K$ diagonal matrix with elements

$$d_k = \sum_{j=1}^J \frac{n_{j+}}{n_{j+} - 1} \frac{n_{j+}}{n_{jk}} \sum_{i=1}^{n_{j+}} \left(\frac{y_{ik} - \mathbf{b}'_k \mathbf{x}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{y_{i'k} - \mathbf{b}'_k \mathbf{x}_{i'}}{N \pi_{i'}} \right)^2.$$

According to (23) it follows that

$$\begin{aligned} \text{Cov}_e(\hat{\mathbf{C}}_{\text{HT}} | m, s) \\ = \mathbf{C} \mathbf{D} \mathbf{C}' - \sum_{j=1}^J \frac{n_{j+}}{n_{j+} - 1} \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{C} \mathbf{E}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{C} \mathbf{E}_{i'}}{N \pi_{i'}} \right) \\ \left(\frac{\mathbf{C} \mathbf{E}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{C} \mathbf{E}_{i'}}{N \pi_{i'}} \right)'. \end{aligned} \quad (51)$$

The final part of the proof is to take the expectation of $\text{Cov}_e(\hat{\mathbf{C}}_{\text{HT}} | m, s)$ with respect to the sampling design and the measurement error model. The proof is given for RBD's where PSU's are block variables. In a two-stage sampling scheme, J blocks or PSU's are drawn from a finite population of J_u blocks with first order inclusion probabilities π_j^I . Within each PSU, n_{j+} SSU's are drawn in the second stage with first and second order inclusion probabilities π_{dj}'' and $\pi_{i|j}''$. The first order inclusion probabilities of the individuals in the sample are $\pi_i = \pi_j^I \pi_{dj}''$. Furthermore, let

$$\bar{\Delta}_j = \sum_{i=1}^{N_j} \frac{\mathbf{e}_i}{N_j}$$

denote the population mean of the measurement errors of the individuals of block j . Then

$$\hat{\Delta}_j = \sum_{i=1}^{n_{j+}} \frac{\mathbf{e}_i}{N_j \pi_{dj}''}$$

denotes the Horvitz-Thompson estimator for $\bar{\Delta}_j$. Now we have

$$\begin{aligned} \sum_{j=1}^J \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{e}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{e}_{i'}}{N \pi_{i'}} \right) \left(\frac{\mathbf{e}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{e}_{i'}}{N \pi_{i'}} \right)' \\ = \sum_{j=1}^J \left(\frac{1}{\pi_j^I} \right)^2 \left(\frac{N_j}{N} \right)^2 \\ \left(\frac{1}{n_{j+}^2} \sum_{i=1}^{n_{j+}} \left(\frac{n_{j+} \mathbf{e}_i}{N_j \pi_{dj}''} - \bar{\Delta}_j \right) \left(\frac{n_{j+} \mathbf{e}_i}{N_j \pi_{dj}''} - \bar{\Delta}_j \right)' \right. \\ \left. - \frac{1}{n_{j+}} (\hat{\Delta}_j - \bar{\Delta}_j) (\hat{\Delta}_j - \bar{\Delta}_j)' \right). \end{aligned} \quad (52)$$

Let E_{s_j} denote the expectation with respect to the first stage of the sampling design and $E_{s_{j+}}$ the expectation with respect to the second stage of the sampling design. Taking the expectation with respect to the measurement error model and the sampling design of the first part of (52) and using model assumption (3) leads to

$$\begin{aligned} E_m E_{s_j} E_{s_{j+}} \sum_{j=1}^J \left(\frac{1}{\pi_j^I} \right)^2 \frac{1}{n_{j+}^2} \sum_{i=1}^{n_{j+}} \left(\frac{n_{j+} \mathbf{e}_i}{N_j \pi_{dj}''} - \bar{\Delta}_j \right) \left(\frac{n_{j+} \mathbf{e}_i}{N_j \pi_{dj}''} - \bar{\Delta}_j \right)' \\ = E_m E_{s_j} \sum_{j=1}^J \left(\frac{1}{\pi_j^I} \right)^2 \frac{1}{n_{j+}} \left(\sum_{i=1}^{N_j} \frac{n_{j+} \mathbf{e}_i \mathbf{e}_i'}{N_j^2 \pi_{dj}''} - \bar{\Delta}_j \bar{\Delta}_j' \right) \\ = \frac{1}{\pi_j^I n_{j+} N_j^2} \sum_{i=1}^{N_j} \left(\frac{n_{j+}}{\pi_{dj}''} - 1 \right) \Sigma_i. \end{aligned} \quad (53)$$

Note that $E_{s_{j+}} (\hat{\Delta}_j - \bar{\Delta}_j) (\hat{\Delta}_j - \bar{\Delta}_j)'$ in (52) equals the design variance of $\hat{\Delta}_j$ with respect to the second stage of the sampling design in block j . Taking the expectation with respect to the measurement error model and the sampling design of the second part of (52) and using model assumption (3) leads to

$$\begin{aligned} E_m E_{s_j} E_{s_{j+}} \sum_{j=1}^J \left(\frac{1}{\pi_j^I} \right)^2 \frac{1}{n_{j+}} (\hat{\Delta}_j - \bar{\Delta}_j) (\hat{\Delta}_j - \bar{\Delta}_j)' \\ = E_m \frac{1}{\pi_j^I N_j^2} \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} (\pi_{i|i}'' - \pi_{dj}'' \pi_{i'j}'') \frac{\mathbf{e}_i \mathbf{e}_{i'}'}{\pi_{dj}'' \pi_{i'j}''} \\ = \frac{1}{\pi_j^I N_j^2} \sum_{i=1}^{N_j} \left(\frac{1}{\pi_{dj}''} - 1 \right) \Sigma_i. \end{aligned} \quad (54)$$

With results (52), (53) and (54) we can elaborate the second term on the right hand side of the equal sign of (51) as

$$\begin{aligned} E_m E_s \sum_{j=1}^J \frac{n_{j+}}{n_{j+} - 1} \sum_{i=1}^{n_{j+}} \left(\frac{\mathbf{C} \mathbf{E}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{C} \mathbf{E}_{i'}}{N \pi_{i'}} \right) \\ \left(\frac{\mathbf{C} \mathbf{E}_i}{N \pi_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{n_{j+}} \frac{\mathbf{C} \mathbf{E}_{i'}}{N \pi_{i'}} \right)' = \frac{1}{N^2} \sum_{i=1}^N \frac{\mathbf{C} \Sigma_i \mathbf{C}'}{\pi_i}. \end{aligned} \quad (55)$$

Finally, it follows from (51) and (55) that

$$E_m E_s \text{Cov}_e(\hat{\mathbf{C}}_{\text{HT}} | m, s) = E_m E_s \mathbf{C} \mathbf{D} \mathbf{C}' - \frac{1}{N^2} \sum_{i=1}^N \frac{\mathbf{C} \Sigma_i \mathbf{C}'}{\pi_i}, \quad \text{QED.}$$

The derivation for an RBD where strata are block variables follows directly as a special case from an RBD where PSU's are block variables with $\pi'_{ij} = 1$, $\pi''_{ij} = \pi_i$, $\pi''_{i|j} = \pi'_i$ and $J = J_u$. The proof for an RBD where clusters are block variables follows directly as a special case from an RBD where PSU's are block variables with $\pi''_{ij} = 1$ and $\pi''_{i|j} = 1$.

The expectation of $\text{Cov}_e(\hat{\mathbf{C}}_{\text{HT}} | m, s)$ with respect to the sampling design and the measurement error model for an RBD where interviewers are the block variables does not follow as a special case from an RBD where PSU's are block variables. Since the block variables are not directly linked with the sampling design, the blocks should be considered as domains where the block size n_{j+} is random with respect to the sampling design. The derivation follows the same steps as in the proof for blocking on PSU's and is given by Van den Brakel (2001).

Proof of formula (32)

Matrix $\hat{\mathbf{D}}$ can be partitioned as follows:

$$\hat{\mathbf{D}} = \begin{pmatrix} \hat{d}_1 & \mathbf{0}' \\ \mathbf{0} & \hat{\mathbf{D}}_* \end{pmatrix}.$$

According to Bartlett's identity (Morisson 1990, chapter 2) it follows that:

$$(\mathbf{C} \hat{\mathbf{D}} \mathbf{C}')^{-1} = (\hat{d}_1 \mathbf{j} \mathbf{j}' + \hat{\mathbf{D}}_*)^{-1} = \hat{\mathbf{D}}_*^{-1} - \frac{1}{\text{trace}(\hat{\mathbf{D}}^{-1})} \hat{\mathbf{D}}_*^{-1} \mathbf{j} \mathbf{j}' \hat{\mathbf{D}}_*^{-1}.$$

From this result it follows that

$$\begin{aligned} \mathbf{C}' (\mathbf{C} \hat{\mathbf{D}} \mathbf{C}')^{-1} \mathbf{C} &= \mathbf{C}' \hat{\mathbf{D}}_*^{-1} \mathbf{C} - \frac{1}{\text{trace}(\hat{\mathbf{D}}^{-1})} \mathbf{C}' \hat{\mathbf{D}}_*^{-1} \mathbf{j} \mathbf{j}' \hat{\mathbf{D}}_*^{-1} \mathbf{C} \\ &= \hat{\mathbf{D}}^{-1} - \frac{1}{\text{trace}(\hat{\mathbf{D}}^{-1})} \hat{\mathbf{D}}^{-1} \mathbf{j} \mathbf{j}' \hat{\mathbf{D}}^{-1}. \end{aligned} \quad (56)$$

Inserting (56) into (31) leads to (32), QED.

Acknowledgements

The authors wish to thank the Associate Editor, the referees, Paul Smith, and Rachel Vis-Visschers for their constructive comments on former drafts of this paper. Jan also thanks Prof. Stephen E. Fienberg and Prof. Peter Kooiman for their support as Ph. D. advisor during this research.

References

- Bethlehem, J.G., and Keller, W.G. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3(2), 141-153.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Fienberg, S.E., and Tanur, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review*, 55(1), 75-96.
- Fienberg, S.E., and Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *The Canadian Journal of Statistics*, 16(2), 135-151.
- Fienberg, S.E., and Tanur, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 243, 1017-1022.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361-374.
- Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart and Winston. 236.
- Hartley, H.O., and Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement*, (Eds. N.K. Namboodiri). New York: Academic Press. 35-43.
- Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York: McGraw-Hill.
- Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian statistical institute. *Journal of the Royal Statistical Society*, 109, 325-370.
- Montgomery, D.C. (2001). *Design and Analysis of Experiments*. New York: John Wiley & Sons, Inc.
- Morisson, D.F. (1990). *Multivariate Statistical Methods*. Singapore: McGraw-Hill.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons, Inc.
- Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.
- Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: Wiley & Sons, Inc. 59-87.
- Statacorp. (2001). *Stata Reference Manual Release 7.0*. College Station, Texas.
- Van den Brakel, J.A. (2001). Design and Analysis of Experiments Embedded in Complex Sample Surveys. Ph.D. Thesis. Rotterdam: Erasmus University of Rotterdam.
- Van den Brakel, J.A. and Binder, D. (2000). Variance estimation for experiments embedded in complex sampling schemes. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Indianapolis, August 13-17. 805-810.
- Van den Brakel, J.A., and Binder, D. (2004). Variance estimation for experiments embedded in complex sampling designs. Unpublished research paper, BPA nr.: H894-04-TMO. Heerlen: Statistics Netherlands.
- Van den Brakel, J.A., and Renssen, R.H. (1998). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14(3), 277-295.
- Van den Brakel, J.A., and Van Berkel, C.A.M. (2002). A design-based analysis procedure for two-treatment experiments embedded in sample surveys. An application in the Dutch labor force survey. *Journal of Official Statistics*, 18(2), 217-231.

Domain Estimators for the Item Count Technique

Takahiro Tsuchiya¹

Abstract

The item count technique, which is an indirect questioning technique, was devised to estimate the proportion of people for whom a sensitive key item holds true. This is achieved by having respondents report the number of descriptive phrases, from a list of several phrases, that they believe apply to themselves. The list for half the sample includes the key item, and the list for the other half does not include the key item. The difference in mean number of selected phrases is an estimator of the proportion. In this article, we propose two new methods, referred to as the cross-based method and the double cross-based method, by which proportions in subgroups or domains are estimated based on the data obtained via the item count technique. In order to assess the precision of the proposed methods, we conducted simulation experiments using data obtained from a survey of the Japanese national character. The results illustrate that the double cross-based method is much more accurate than the traditional stratified method, and is less likely to produce illogical estimates.

Key Words: Indirect questioning techniques; Item count technique; Domain estimators; Survey of Japanese national character.

1. Introduction

1.1 Indirect Questioning Techniques

Suppose that a population U is divided into two subpopulations $U_{(T)}$ and $U_{(T)}^c$, where $U_{(T)}$ is a set of elements having an attribute T , and $U_{(T)}^c$ is a complement of $U_{(T)}$. One purpose of social surveys is to estimate $\pi = \bar{Y} = P(Y=1)$, where

$$Y_k = \begin{cases} 1 & \text{if } k \in U_{(T)} \\ 0 & \text{otherwise} \end{cases}$$

and $P(\cdot)$ denotes the proportion of units having a particular value of the variable. For example, when T is “supporting the present cabinet,” π indicates the cabinet support rate, and when T is “using a certain illegal drug,” π denotes the prevalence rate of drug use.

In a direct questioning technique, researchers ask respondents “Do you belong to $U_{(T)}$?” and directly obtain the indicator value y_i as “yes” or “no” (Cochran 1977, page 50). When every respondent has an equal inclusion probability, a sample mean \bar{y} serves as one estimator of π .

On the other hand, some indirect questioning techniques, including the randomized response technique (Warner 1965), the nominative technique (Miller 1985), the item count technique (Droitcour, Caspar, Hubbard, Parsley, Visscher and Ezzati 1991), and the three-card technique (Droitcour, Larson and Scheuren 2001), are devised because some respondents tend to evade sensitive questions, such as those concerning highly private matters, socially unacceptable or deviant behaviors or illegal acts. The essential feature of

indirect techniques is that instead of a direct observation of Y , another variable $X = g(Y, V)$, which is some sort of function of Y and, if necessary, of other random variables V , is observed so that respondents feel that their true Y -values are not revealed. While this feature is expected to derive a truthful answer from evasive respondents, both the questioning and the estimation procedures are rather complicated compared to the direct questioning technique partly because the function $g(\cdot)$ sometimes includes some randomization processes. We shall outline two indirect techniques below.

The randomized response is the most popular among the indirect techniques, and various modifications have been proposed (Abul-Elä, Greenberg and Horvitz 1967; Warner 1971; Chaudhuri and Mukerjee 1988; Greenberg, Abul-Elä, Simmons and Horvitz 1969; Takahasi and Sakasegawa 1977). Although the randomized response is not the topic of this article, we shall briefly outline Warner’s original procedure here for reference, because this technique will be simulated in a later section.

1. Prepare two types of questionnaires. In questionnaire A , respondents are asked “Do you belong to $U_{(T)}$?” and in questionnaire B , respondents are asked “Do you belong to $U_{(T)}^c$?”
2. Let $p (\neq 0.5)$ be the predetermined probability. Each respondent selects questionnaire A or B with probabilities p or $1-p$ respectively, but no one other than the respondent knows which questionnaire is selected.

1. Takahiro Tsuchiya, The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan. E-mail: taka@ism.ac.jp.

3. Suppose X is an indicator variable whose value is 1 if the response is "yes" or 0 if the response is "no." The estimator of π is given by

$$\hat{\pi} = \frac{p - 1 + \bar{x}}{2p - 1}, \quad (1)$$

where \bar{x} is a sample mean of X .

Since the researchers have no information regarding the type of questionnaire selected by each respondent, more respondents are expected to give truthful answers than they would if asked direct questions.

The item count technique, which is the subject of this article, is not as popular despite its simplicity. The technique is also effective when posing sensitive questions, because respondents are asked not to answer sensitive questions directly but to merely report the number of items that hold true with them. The following are the processes of the item count technique:

1. Prepare the key item T , which is the primary focus of the study, and G other non-key items E_1, \dots, E_G . For example, T is "using a certain illegal drug" as mentioned above, and E_g is some sort of non-sensitive description such as "owning a bicycle."
2. Prepare two types of questionnaires, A and B . In questionnaire A , respondents are asked to answer the number C^A of items that are true with respect to themselves among G non-key items. In questionnaire B , respondents are asked to answer the number C^B of items that are true with respect to themselves out of $G+1$ items, including the key item T .

Table 1 lists examples of item lists. Our aim is to estimate the proportion of people who use a certain illegal drug. The key item is "using a certain illegal drug" in the questionnaire B and the other four items are non-key items. Except when a response to the questionnaire B is $C^B = 0$ or $C^B = 5$, researchers cannot detect as to which items hold true with the respondent. For example, a respondent will reply that four items in the questionnaire B are true, but we cannot be sure that the respondent uses the drug at all. Hence, it is expected that more respondents using an illegal drug will report truthful answers in such a scenario than when asked a direct question.

3. Divide a total sample into two subgroups, A and B , randomly of size n^A and n^B so that each questionnaire is assigned to a corresponding subgroup.

Table 1
Examples of Item Lists

Questionnaire A	Questionnaire B
How many of the following hold true for you?	How many of the following hold true for you?
– owning a bicycle	– owning a bicycle
– having travelled abroad	– having travelled abroad
– having called an ambulance	– having called an ambulance
– owning a summer villa	– using a certain illegal drug
	– owning a summer villa

4. The estimator of π is given by

$$\hat{\pi} = \hat{C}^B - \hat{C}^A, \quad (2)$$

where \hat{C}^A and \hat{C}^B are the estimated means of C^A and C^B respectively. The justification of (2) is explained in section 2.1. When every unit in the sample has an equal inclusion probability, $\hat{\pi}$ can be written as

$$\hat{\pi} = \sum_{c=0}^{G+1} c \frac{n_c^B}{n^B} - \sum_{c=0}^G c \frac{n_c^A}{n^A}, \quad (3)$$

where n_c^A and n_c^B are the number of respondents whose answers are $C^A = c$ and $C^B = c$, respectively. Moreover, when an auxiliary variable Z is available and its distribution $P(Z = z) = m_z$ in the population is known, for example from a census, poststratification is often used to adjust the sample distribution of Z to the population. That is, the poststratified estimator of π is given by

$$\begin{aligned} \hat{\pi}_{PS} &= \sum_{c=0}^{G+1} c \frac{\sum_z v_z^B n_{cz}^B}{n^B} - \sum_{c=0}^G c \frac{\sum_z v_z^A n_{cz}^A}{n^A} \\ &= \sum_{c=0}^{G+1} c \sum_z \frac{m_z}{n_z} \frac{n_{cz}^B}{n_z} - \sum_{c=0}^G c \sum_z \frac{m_z}{n_z} \frac{n_{cz}^A}{n_z}, \end{aligned} \quad (4)$$

where n_{cz}^A is the number of respondents for each $C^A = c$ and $Z = z$,

$$n_z^A = \sum_{c=0}^G n_{cz}^A, \quad n^A = \sum_z n_z^A, \quad v_z^A = \frac{m_z n^A}{n_z^A}$$

and n_{cz}^B , n_z^B , n^B , and v_z^B are defined in analogous ways.

One practical merit of the item count technique is that it does not demand any randomization devices, which are required for the randomized response technique. It is not the respondent but a researcher who selects the questionnaire to be answered. Hence, the item count technique is easily implemented via any self-administered or telephone surveys. A more elaborate comparison between the randomized response and the item count technique is found in Hubbard, Casper and Lessler (1989).

The questionnaire A is introduced to obtain the distribution of the number of non-key items. That is, respondents to the questionnaire A do not answer the sensitive question. Therefore, it is possible to increase the precision of the estimator using the double-list version of item count (Droitcour *et al.* 1991), which exchanges the roles between the two subgroups. However, we limit our argument in this article to a single-list version, because the extension of estimators to the double-list version is straightforward.

1.2 Purpose of this Article

Thus far, we have focused on the parameter $\pi = \bar{Y} = P(Y=1)$ of a total population. However, estimators in subpopulations or domains (Särndal, Swesson and Wretman 1992 page 5) are often required, *i.e.*, either a conditional proportion $P(Y=1|Z=z)$ or a joint proportion $P(Y=1, Z=z)$ must be estimated, where a population is divided into several domains by the Z -value. We refer to the variable Z as the domain variable in this article. The domain variables often used are demographic characteristics such as gender or age. For example, government agencies would like to know the proportion of people who use a certain illegal drug at each age group. Even though the post-stratified estimator $\hat{\pi}_{ps}$ in (4) uses the domain variable Z , its aim is an estimation of $P(Y=1)$ in the entire population. Our aim in this article is to obtain separate estimations of $P(Y=1|Z=z)$ within each domain.

One simple estimation method is as follows:

1. Post-stratify the sample into strata or domains based on the Z -value.
2. In each stratum or domain, separately determine $p(Y=1|Z=z)$ using (1) or (2), where $p(\cdot)$ is a sample estimate of $P(\cdot)$.
3. If necessary, estimate $p(Y=1, Z=z)$ by multiplying a known domain proportion, $P(Z=z)$, or an estimated domain proportion, $p(Z=z)$.

The above method is referred to throughout this article as a stratified method because estimates are obtained separately in each stratum or domain. Although Rao (2003) refers to the above method as a direct estimate, we have avoided the use of the term "direct" in order to avoid confusion with the term "direct questioning technique."

An advantage of the stratified method is that this method is applicable to any indirect questioning technique, including the randomized response and item count techniques. The U.S. General Accounting Office (1999) adopts the stratified method to estimate domains under the three-card technique. However, one of the serious problems of the stratified method is that it often produces illogical estimates, especially negative estimates, in the case of the randomized response and the item count, as explained later

in this article. This is mainly because the reduction of the sample size in each stratum increases the standard errors of the estimators (Lessler and O'Reilly 1997). For example, Droitcour *et al.* (1991, page 206) "calculated estimates separately for the three risk strata" and obtained negative prevalence rate estimates of drug use.

In the case of the randomized response, there is little possibility that domain estimators other than the stratified method are developed because information concerning the type of questionnaire selected by individual respondents is unavailable. In contrast, in the item count technique, the questionnaire answered by each respondent is known. Therefore, the precision of the domain estimators is expected to increase when auxiliary information is used, specifically contingency tables between Z and C^A or C^B .

In this article, we propose new domain estimators for the item count technique, which are referred to as the cross-based method and the double cross-based method. In addition, we will illustrate the fact that the new estimators are more efficient than the traditional stratified method by simulating the item count technique using data obtained from the survey of the Japanese national character concerning the significant attributes of the Japanese character.

2. Domain Estimators for the Item Count Technique

2.1 Stratified Method

Here, we reformulate the stratified method. Let us assume that the following equations hold true for each value of c and z .

Assumption 1.

$$\begin{aligned} P(C^B = c|Z = z) &= P(C^A = c, Y = 0|Z = z) \\ &\quad + P(C^A = c - 1, Y = 1|Z = z), \\ P(C^A = G + 1, Y = 0|Z = z) &= 0. \end{aligned}$$

These assumptions imply that the difference in the distribution between C^A and C^B depends solely on Y . Question effects, including order effects and context effects (Schuman and Presser 1981) are not considered.

We have the following result based on these assumptions.

Stratified Method.

$$\begin{aligned} P(Y = 1|Z = z) &= \sum_{c=0}^{G+1} c P(C^B = c|Z = z) \\ &\quad - \sum_{c=0}^G c P(C^A = c|Z = z) \quad (5) \\ &= \bar{C}_z^B - \bar{C}_z^A, \quad (6) \end{aligned}$$

where \bar{C}_z^A and \bar{C}_z^B are the domain means of C^A and C^B .

Derivation.

$$\begin{aligned} & \sum_{c=0}^{G+1} cP(C^B = c|Z = z) \\ &= \sum_{c=0}^{G+1} cP(C^A = c, Y=0|Z=z) + \sum_{c=0}^{G+1} cP(C^A = c-1, Y=1|Z=z) \\ &= \sum_{c=0}^G cP(C^A = c, Y=0|Z=z) + \sum_{c=0}^G (c+1)P(C^A = c, Y=1|Z=z) \\ &= \sum_{c=0}^G c\{P(C^A = c, Y=0|Z=z) + P(C^A = c, Y=1|Z=z)\} \\ &\quad + \sum_{c=0}^G P(C^A = c, Y=1|Z=z) \\ &= \sum_{c=0}^G cP(C^A = c|Z=z) + P(Y=1|Z=z). \end{aligned}$$

Transposing the first term to the left-hand side yields the stratified method (5).

The estimator $p(Y=1|Z=z)$ is obtained by substituting domain means \bar{C}_z^A and \bar{C}_z^B with their estimators, $\hat{\bar{C}}_z^A$ and $\hat{\bar{C}}_z^B$.

$$p(Y=1|Z=z) = \hat{\bar{C}}_z^B - \hat{\bar{C}}_z^A. \quad (7)$$

When the inclusion probabilities are equal for all units in the sample, the estimator of $P(Y=1|Z=z)$ is written as

$$p(Y=1|Z=z) = \sum_{c=0}^{G+1} c \frac{n_{cz}^B}{n_z^B} - \sum_{c=0}^G c \frac{n_{cz}^A}{n_z^A}, \quad (8)$$

where n_{cz}^A , n_{cz}^B , n_z^A , and n_z^B are defined in the section 1.1. The equations (2) and (3) for the entire population are special cases of (7) and (8).

One merit of the stratified method is that the variance estimator of $p(Y=1|Z=z)$ is easily obtained by

$$\hat{\text{Var}}(p(Y=1|Z=z)) = \hat{\text{Var}}(\hat{\bar{C}}_z^B) + \hat{\text{Var}}(\hat{\bar{C}}_z^A). \quad (9)$$

On the other hand, as noted in the previous section, the reduction of sample size in each stratum increases estimated variances in (9). Further, the marginal estimator $p(Y=1)$ obtained by using (8) does not correspond to that obtained directly by (3), unless $n_z^A = n_z^B$ for all z . That is, when $p(Z=z)$ is not known, its estimator is given by

$$p(Z=z) = (n_z^A + n_z^B) / (n^A + n^B)$$

and

$$\begin{aligned} & \sum_z p(Y=1|Z=z)p(Z=z) \\ &= \sum_z \frac{n_z^A + n_z^B}{n^A + n^B} \left\{ \sum_{c=0}^{G+1} c \frac{n_{cz}^B}{n_z^B} - \sum_{c=0}^G c \frac{n_{cz}^A}{n_z^A} \right\} \\ &\neq \sum_{c=0}^{G+1} c \frac{n_c^B}{n^B} - \sum_{c=0}^G c \frac{n_c^A}{n^A} = \hat{\pi}. \end{aligned} \quad (10)$$

When the domain proportion $p(Z=z) = m_z$ is available, the marginal estimator corresponds to the poststratified estimator (4).

$$\begin{aligned} & \sum_z p(Y=1|Z=z)P(Z=z) \\ &= \sum_z m_z \left\{ \sum_{c=0}^{G+1} c \frac{n_{cz}^B}{n_z^B} - \sum_{c=0}^G c \frac{n_{cz}^A}{n_z^A} \right\} \\ &= \hat{\pi}_{\text{PS}}. \end{aligned}$$

These results indicate that we should use a poststratified estimator $\hat{\pi}_{\text{PS}}$ with the domain estimators if we use the stratified method.

2.2 Cross-based Method

In the stratified method, a total sample is divided into strata for the purpose of direct estimation of $P(Y=1|Z=z)$, which causes sample size reduction. Hence, in the cross-based method proposed in this section, the joint proportion $P(Y=1, Z=z)$ is estimated first in order to use the entire sample, and the conditional proportion is subsequently obtained by

$$p(Y=1|Z=z) = \frac{p(Y=1, Z=z)}{p(Z=z)}$$

$$\text{or } p(Y=1|Z=z) = \frac{p(Y=1, Z=z)}{P(Z=z)}.$$

The term ‘cross-based method’ is used because this method uses cross tabulations $P(Z=z|C^B=c)$, as shown in (19).

For the cross-based method, we assume that the following equations hold for each value of c .

Assumption 2.

$$P(C^B = c+1, Y=1) = P(C^A = c, Y=1), \quad (11)$$

$$P(C^B = 0, Y=1) = P(C^A = -1, Y=1) = 0, \quad (12)$$

$$P(C^B = c, Y=0) = P(C^A = c, Y=0). \quad (13)$$

These assumptions also imply that the difference in the distribution between C^A and C^B depends only on Y .

We have the following result based on these assumptions.

Cross-based Method.

$$P(Y=1, Z=z) = \sum_{c=1}^{G+1} P(Z=z|C^B=c)Q_{c-1}, \quad (14)$$

where

$$Q_c = \sum_{d=0}^c \{P(C^A = d) - P(C^B = d)\}.$$

In addition, we assume that $P(Z=z|C^B=c, Y=1) = P(Z=z|C^B=c)$ for every $c > 0$. This assumption would be valid to some degree when both the key and non-key items describe the same type of stigmatizing behavior.

Derivation.

Based on the assumptions, we have

$$\begin{aligned} P(C^B = c) &= P(C^B = c, Y=1) + P(C^B = c, Y=0) \\ &= P(C^A = c-1, Y=1) + P(C^A = c, Y=0). \end{aligned} \quad (15)$$

The following equation holds for any c .

$$P(C^A = c, Y=0) = P(C^A = c) - P(C^A = c, Y=1). \quad (16)$$

Hence, substituting (16) in (15) gives

$$\begin{aligned} P(C^B = c) &= P(C^A = c-1, Y=1) \\ &\quad + \{P(C^A = c) - P(C^A = c, Y=1)\}. \end{aligned} \quad (17)$$

Summing (17) over c up to some g , we obtain

$$\begin{aligned} \sum_{c=0}^g P(C^B = c) &= \sum_{c=0}^g P(C^A = c-1, Y=1) \\ &\quad + \sum_{c=0}^g \{P(C^A = c) - P(C^A = c, Y=1)\} \\ &= \sum_{c=0}^g P(C^A = c) - P(C^A = g, Y=1). \end{aligned}$$

By transposing the terms, we define Q_c .

$$\begin{aligned} Q_c &= \sum_{d=0}^c \{P(C^A = d) - P(C^B = d)\} \\ &= P(C^A = c, Y=1) \\ &= P(C^B = c+1, Y=1). \end{aligned} \quad (18)$$

Here, the joint proportion $P(Y=1, Z=z)$ is decomposed as

$$P(Y=1, Z=z) = \sum_{c=0}^{G+1} P(Z=z|C^B=c)P(C^B=c, Y=1). \quad (19)$$

Substituting the equation (18) and the assumption (12) in (19) yields the cross-based method.

The joint estimator $P(Y=1, Z=z)$ is obtained by substituting each term of (14) for its estimators. When the sample is self-weighting, the estimator is given by

$$P(Y=1, Z=z) = \sum_{c=1}^{G+1} \frac{n_{cz}^B}{n_c^B} \sum_{d=0}^{c-1} \left(\frac{n_{dz}^A}{n^A} - \frac{n_{dz}^B}{n^B} \right), \quad (20)$$

where

$$n_c^A = \sum_z n_{cz}^A \quad \text{and} \quad n_c^B = \sum_z n_{cz}^B.$$

The conditional estimator $p(Y=1|Z=z)$ is obtained by dividing $p(Y=1|Z=z)$ by the domain proportions $P(Z=z)$ or their estimators $p(Z=z)$.

As noted above, the main feature of the cross-based method is that $p(Y=1, Z=z)$ is first estimated using the

entire sample. Hence, the variance of $p(Y=1|Z=z)$ for the cross-based method is expected to be smaller than that of $p(Y=1|Z=z)$ for the stratified method. Moreover, negative values will seldom be obtained in the case of the cross-based method, while the negative values will be often obtained in the case of the stratified method. Furthermore, the marginal estimator $p(Y=1)$ obtained by summing (20) is equal to the estimator (3), unless $n_c^B = 0$ for some c :

$$\begin{aligned} \sum_z p(Y=1, Z=z) &= \sum_z \sum_{c=1}^{G+1} \frac{n_{cz}^B}{n_c^B} \sum_{d=0}^{c-1} \left(\frac{n_{dz}^A}{n^A} - \frac{n_{dz}^B}{n^B} \right) \\ &= \sum_{c=1}^{G+1} \sum_{d=0}^{c-1} \left(\frac{n_{d\cdot}^A}{n^A} - \frac{n_{d\cdot}^B}{n^B} \right) \\ &= \sum_{c=1}^{G+1} \left\{ \left(1 - \sum_{d=c}^G \frac{n_{d\cdot}^A}{n^A} \right) - \left(1 - \sum_{d=c}^{G+1} \frac{n_{d\cdot}^B}{n^B} \right) \right\} \\ &= \sum_{c=0}^{G+1} c \frac{n_c^B}{n^B} - \sum_{c=0}^G c \frac{n_c^A}{n^A} = \hat{\pi}. \end{aligned} \quad (21)$$

Of course, when the domain proportions $P(Z=z) = m_z$ are known, we can use them to obtain a poststratified estimator $p(C^A = d)$ of $P(C^A = d)$ in Q_{c-1} of (14),

$$p(C^A = d) = \sum_z \frac{m_z}{n_z^B} n_{dz}^B.$$

In this case, $\sum_z p(Y=1, Z=z)$ coincides with the post-stratified estimator $\hat{\pi}_{PS}$.

One drawback of the cross-based method is that the variance of $p(Y=1|Z=z)$ is almost impossible to estimate algebraically. Hence, some resampling methods such as the jackknife or bootstrap would be necessary. Additionally, since it is impossible to determine the more efficient method between the stratified method and the cross-based method, simulation studies shall be conducted in a later section.

2.3 Double Cross-based Method

Before proceeding to the simulation study, we suggest a modified version of the cross-based method. In equation (19) of the cross-based method, we use $P(Z=z|C^B=c)$. In the same way, when $P(Z=z|C^A=c)$ is used, we obtain

$$\begin{aligned} P(Y=1, Z=z) &= \sum_{c=0}^G P(Z=z|C^A=c)P(C^A=c, Y=1) \\ &= \sum_{c=0}^G P(Z=z|C^A=c)Q_c. \end{aligned} \quad (22)$$

Hence, a double cross-based method is obtained by combining (14) and (22) as follows:

$$P(Y=1, Z=z) = \sum_{c=0}^G \left\{ w^A P(Z=z|C^A=c) + w^B P(Z=z|C^B=c+1) \right\} Q_c, \quad (23)$$

where w^A and w^B are the non-negative weights for each subgroup, the sum of which is equal to one.

The following equation also holds for the double cross-based method of any w^A and w^B , unless $n_c^A = 0$ or $n_c^B = 0$ for some c .

$$\sum_z p(Y = 1, Z = z) = \hat{\pi}. \tag{24}$$

3. Numerical Experiments

3.1 Data Set

In order to compare the precision of the estimators, we conducted simulation experiments using data obtained from the survey of the Japanese national character (Sakamoto, Tsuchiya, Nakamura, Maeda and Fouse 2000). Although the respondents were selected via a stratified two-stage sampling from Japanese aged 20 and over, we neglect the sampling design because the collected sample of $N = 1,339$ is treated as the “true” population in this experiment. Table 2 lists the results of a question concerning the significant attributes of the Japanese character. Respondents were asked in a face-to-face interview to choose as many adjectives from among ten alternatives as they thought described the Japanese character.

Table 2
Significant Attributes of Japanese character

$N = 1,339$				
(Hand card) Which of the following adjectives do you think describes the character of the Japanese people? Choose as many as you like.				
1 Rational	18%	6 Kind	42%	
2 Diligent	71%	7 Original	7%	
3 Free	13%	8 Polite	50%	
4 Open, frank	14%	9 Cheerful	8%	
5 Persistent	51%	10 Idealistic	23%	

The form of this question is different from that of the item count technique. In the item count technique, the respondent is asked to “answer the number of adjectives.” In contrast, in this survey the respondent is asked to “circle as many adjectives you feel are appropriate.” In addition, the ten items are not very sensitive, hence the respondents should not hesitate during the selection. However, since the real contingency table between each of the ten items and another variable Z is obtained, we can evaluate the performance of estimators through a pseudo item count procedure.

We took each of the following three items as the key item Y , where $Y = 1$ implies that the item was selected.

- 7 Original (π is the least among the ten items)
- 8 Polite (π is just 50%)
- 2 Diligent (π is the largest among the ten items)

Three combinations of non-key items are used, as listed in Table 3. Combination 1 comprises two items with low proportions, while combination 2 comprises two items with high proportions. Combination 3 is the case with the maximum number of non-key items.

Table 3
Three Combinations of Non-key Items

	Non-key items	
Combination 1 ($G = 2$):	9 Cheerful	(8%)
	3 Free	(13%)
Combination 2 ($G = 2$):	5 Persistent	(51%)
	6 Kind	(42%)
Combination 3 ($G = 9$):	Nine items other than the key item	

We used either gender or age as the domain variable Z . Gender is either male or female, and the age categories are “20 – 29,” “30 – 39,” “40 – 49,” “50 – 59,” “60 – 69”, and “70 and over.”

3.2 Direct Questioning Versus Item Count Technique

3.2.1 Simulation Methods

First, we compare the standard errors between the direct questioning and the item count techniques. In this experiment, we attempted one combination of “7 Original” (key item), combination 3 (non-key items), and gender (domain variable). The contingency table based on the entire sample of $N = 1,339$ is listed in Table 4.

Table 4
A Contingency Table Between “7 Original” and Gender

	7 Original		Total	
	$Y = 1$	$Y = 0$		
Male	46 (7.5)	569 (92.5)	615	(100.0)
Female	51 (7.0)	673 (93.0)	724	(100.0)
Total	97 (7.2)	1,242 (92.8)	1,339	(100.0)

The simulation was conducted through the following procedures:

- Step 1. Suppose the total sample of $N = 1,339$ to be a population.
- Step 2. Draw a subsample S of size Nf where f is a sampling fraction with a simple random sampling without replacement.
- Step 3. As the simulated result of the direct questioning method, compute the proportion directly, $p(Y = 1|Z = \text{male})$ and $p(Y = 1|Z = \text{female})$.
- Step 4. Divide the subsample S into two groups S^A and S^B of size n^A and n^B that are not necessarily of equal size. Count the number C^A of selected non-key items for each respondent in S^A . Also, count the number C^B of selected items including both the key item and the non-key items in S^B .

- Step 5. As the simulated result of the item count technique, compute $p(Y=1|Z=\text{male})$ $p(Y=1|Z=\text{female})$ and via the three estimation methods; stratified method, cross-based method, and double cross-based method. In the double cross-based method, we let $w^A = n^A / (n^A + n^B)$ and $w^B = n^B / (n^A + n^B)$.
- Step 6. We let $f = 0.1$ in step 2 and perform steps 2 to 5 for 2,000 iterations. Calculate the means E_D, E_S, E_C , and E_W and the standard deviations SE_D, SE_S, SE_C , and SE_W of each estimation method to approximate the expectations and the standard errors of the estimators, where the subscripts D, S, C , and W , indicate the direct questioning method, the stratified method, the cross-based method, and the double cross-based method, respectively. In the same way, we let $f = 0.2$ and perform steps 2 to 5 for 2,000 iterations, and so on up to and including $f = 0.9$.

3.2.2 Simulation Results

Figure 1 shows the approximated expectations and standard errors of the estimators. The horizontal axes indicate sampling fraction f . In both the cases, male and female, the approximated expectations of E_D are stable at every f -value while E_S, E_C , and E_W of the item count technique fluctuate irregularly. This is because randomness is introduced twice under the item count, *i.e.*, in the sampling phase and in the division phase, whereas randomness is introduced only in the sampling phase under the direct questioning scenario. Even if $f = 1$, the estimator under the item count technique has a certain amount of variance due to the randomness at the division phase. As the range of fluctuation was negligible compared to the magnitude of the standard errors, which are referred to below, we concluded that the number of repetition was sufficient.

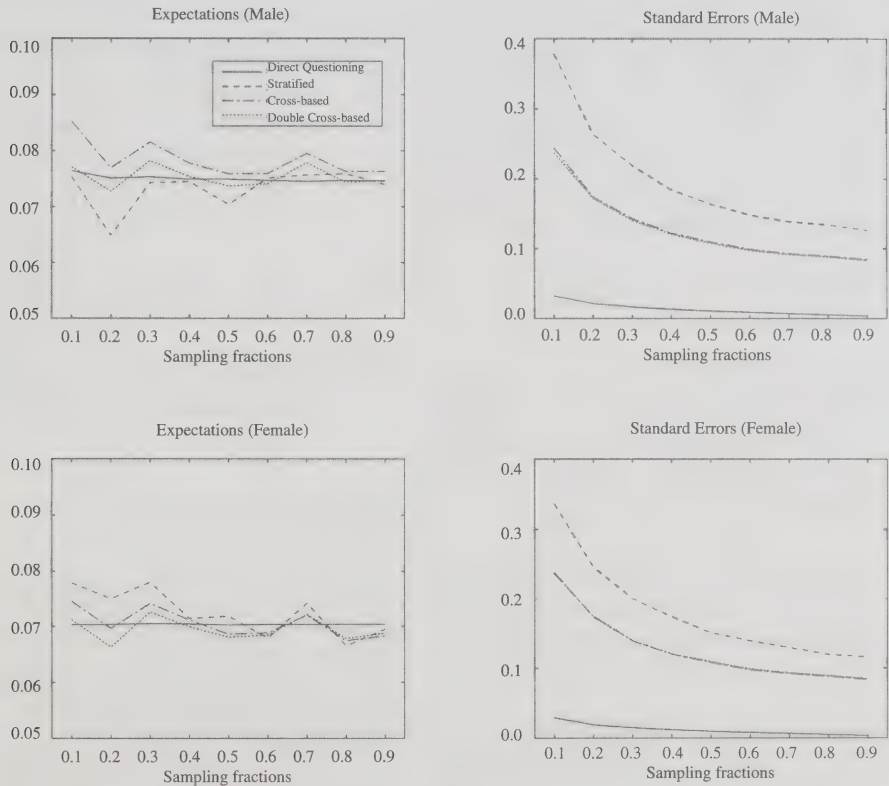


Figure 1. Approximated Expectations and Standard Errors of Estimators.

The standard errors, SE_D , of the direct questioning method is considerably small compared to those of the item count. In the case of the item count, standard errors do not converge to zero even if $f = 1$. As noted above, this is because the randomness is also introduced in the division phase. The standard errors of the stratified method are obviously larger than those of the two cross-based methods. The lines indicating the results for the cross-based method and the double cross-based method almost overlap, and appear to have no outstanding differences.

In order to evaluate the amount of variances or standard errors of estimators, let us consider the following indices that are analogous to the design effect (Kish 1965),

$$\text{Def}_{M_1, M_2} = \frac{SE_{M_1}^2}{SE_{M_2}^2},$$

where M_1 and M_2 indicate one of the four methods D , S , C , and W . Although we have omitted the detailed results, roughly summarized, $\text{Def}_{C,D}$ ranges from 50 (when $f = 0.1$) to 700 (when $f = 0.9$). That is, even if we use the cross-based method, the standard errors of the item count inflate nearly seven- to twenty-six-fold as compared to the direct questioning. However, the variance reduction attained by using the double cross-based method instead of the stratified method ranges from $\text{Def}_{W,S} = 0.39$ (male) to 0.55 (female). In other words, the standard errors of the double cross-based method are reduced to about 62 percent of the stratified estimate at the minimum, and 74 percent at the maximum.

3.3 Stratified Versus Cross-based Method

3.3.1 Simulation Methods

In the previous section, the precision of the cross-based and the double cross-based method appeared to be larger than those of the stratified method. We shall check the precision of these methods for other combinations of the key item, the combination of non-key items, and the domain variable Z by simulation experiments.

In this section, we used all samples as follows:

- Step 1. Compute $P(Y=1|Z=z)$ for each z based on all data of size $N=1,339$.
- Step 2. Divide the total sample ($N=1,339$) randomly into group A and group B of size n^A and n^B where $N = n^A + n^B$.
- Step 3. Count the number C^A of selected non-key items for each respondent of group A , and count the number C^B of selected items, including both the key item and non-key items, in group B .

- Step 4. Estimate $p(Y=1|Z=z)$ by the stratified method, the cross-based method, and the double cross-based method, respectively.

- Step 5. Compute the chi-squared distance e^2 between $P(Y=1|Z=z)$ and $p(Y=1|Z=z)$ for each method.

$$e^2 = \sum_z \frac{\{p(Y=1|Z=z) - P(Y=1|Z=z)\}^2}{P(Y=1|Z=z)}$$

- Step 6. Repeat the above procedure from step 2 through step 5 for 1,000 iterations. Calculate the means and the standard deviations of e^2 for each method.

In addition, we simulated the stratified method under the randomized response for references via the following procedure:

- Step 1. Let p be a proportion as described below. Divide the total sample ($N=1,339$) randomly into two groups. Group A is composed of Np respondents, and group B is composed of $N(1-p)$ respondents.

- Step 2. Let n_z^A be the number of respondents who selected the key item and $Z=z$ in group A . Let n_z^B be the number of respondents who did not select the key item and $Z=z$ in group B . Let n_z be the number of respondents with $Z=z$. Compute

$$p(Y=1|Z=z) = \frac{n_z}{1,339} \left(\frac{p-1 + (n_z^A + n_z^B)/n_z}{2p-1} \right).$$

- Step 3. Calculate e^2 employing the same equation as used in the item count technique.

- Step 4. Repeat the above procedure from step 1 through step 3 for 1,000 iterations. Calculate the means and the standard deviations of e^2 for each method.

We used three p values; $p=0.2$, $p=0.3$, and $p=0.4$.

3.3.2 Simulation Results

Table 5 and Table 6 list the means and the standard deviations of 1,000 e^2 s for the domain variable Z of gender and age, respectively. A smaller mean of “ e^2 -value” indicates that the domain estimators are more precise. In some repetitions, illogical estimates $p(Y=1|Z=z)$, which deviate from the range $[0, 1]$, were obtained. The columns of the tables denoted by “under” indicate the number of repetitions when at least one of the estimates $p(Y=1|Z=z)$ was under 0, and “over” indicates that the estimates were over 1. Ideally, the figures of the columns of “illogical p ” should be 0.

Table 5

Means and Standard Deviations of e^2 s and Number of Times Illogical Estimates were Obtained (Domain Variable Z is Gender)

	7 Original (7%)				8 Polite (50%)				2 Diligent (71%)			
	e^2 -value mean	(s.d.)	illogical under	p over	e^2 -value mean	(s.d.)	illogical under	p over	e^2 -value mean	(s.d.)	illogical under	p over
Stratified method												
Combination 1	38	(36)	39	0	6	(6)	0	0	4	(4)	0	0
Combination 2	89	(92)	179	0	16	(17)	0	0	10	(11)	0	0
Combination 3	341	(330)	457	0	44	(43)	0	0	33	(32)	0	7
Cross-based method												
Combination 1	18	(24)	1	0	4	(5)	0	0	3	(3)	0	0
Combination 2	45	(65)	41	0	10	(12)	0	0	7	(8)	0	0
Combination 3	163	(239)	186	0	22	(31)	0	0	17	(23)	0	1
Double cross-based method												
Combination 1	18	(24)	1	0	3	(4)	0	0	2	(3)	0	0
Combination 2	45	(65)	31	0	9	(12)	0	0	6	(8)	0	0
Combination 3	163	(240)	177	0	21	(31)	0	0	16	(23)	0	0
Randomized response												
$p = 0.2$	12	(14)	0	0	3	(3)	0	0	2	(2)	0	0
$p = 0.3$	35	(43)	41	0	8	(7)	0	0	5	(5)	0	0
$p = 0.4$	158	(181)	305	0	35	(34)	0	0	23	(23)	0	3

Note: e^2 -value is multiplied by 10^3 .

Table 6

Means and Standard Deviations of e^2 s and Number of Times Illogical Estimates were Obtained (Domain Variable Z is age)

	7 Original (7%)				8 Polite (50%)				2 Diligent (71%)			
	e^2 -value mean	(s.d.)	illogical under	p over	e^2 -value mean	(s.d.)	illogical under	p over	e^2 -value mean	(s.d.)	illogical under	p over
Stratified method												
Combination 1	375	(226)	609	0	60	(39)	0	0	39	(26)	0	0
Combination 2	859	(507)	799	0	152	(91)	0	0	97	(58)	0	18
Combination 3	3,410	(2,108)	926	1	446	(290)	48	41	333	(217)	9	353
Cross-based method												
Combination 1	93	(82)	8	0	32	(20)	0	0	28	(16)	0	0
Combination 2	175	(195)	138	0	80	(42)	0	0	59	(33)	0	0
Combination 3	536	(733)	273	0	89	(95)	0	0	70	(71)	0	10
Double cross-based method												
Combination 1	70	(75)	8	0	13	(13)	0	0	9	(8)	0	0
Combination 2	153	(202)	93	0	45	(35)	0	0	31	(23)	0	0
Combination 3	526	(745)	246	0	72	(94)	0	0	52	(70)	0	1
Randomized response												
$p = 0.2$	158	(101)	284	0	25	(14)	0	0	17	(11)	0	0
$p = 0.3$	476	(294)	720	0	74	(42)	0	0	51	(31)	0	2
$p = 0.4$	2,181	(1,348)	945	0	335	(193)	9	9	232	(136)	0	217

Note: e^2 -value is multiplied by 10^3 .

For every combination of the key item, the non-key items, and the domain variable Z, the means of e^2 of the double cross-based method are the smallest, and the cross-based method is the second smallest by a narrow margin. When π of the key item is low ("7 Original"), the number of non-key items is large (combination 3), and the number of alternatives of the domain variable Z is large (age), the accuracy of the stratified method decreases greatly compared to other combinations.

Moreover, when π of the key item is low, negative estimates are often observed when the stratified method is

used. For example, when combining "7 Original," combination 3 and age, the frequency of observed negative estimates is 926 out of 1,000 iterations. When the double cross-based method is used, the negative estimates are less likely to be observed.

For randomized response, when the number of alternatives of the domain variable Z is small (gender), the accuracy of the estimates seems to be the same as the cross-based and the double cross-based methods. However, the mean e^2 is somewhat larger than that of the cross-based method when the domain variable Z has many options (age).

The randomized response, for which only the stratified method is available, also suffers from negative estimates, particularly when π is small ("7 Original").

4. Conclusion

The following results were obtained through simulation experiments:

- The cross-based method or the double cross-based method, which is proposed in this article, should be used to estimate domain parameters when the data is obtained via the item count technique. In the first simulation, the variances of cross-based estimators were reduced to 39 percent of the variance of the stratified estimate at the minimum to 55 percent at the maximum. In the simulation studies, the double cross-based method made no drastic improvement in precision as compared to the cross-based method.
- Even when the double cross-based method is used, the standard errors of the domain estimators are much larger than those of the direct questioning technique.

The true $\pi = \bar{Y} = P(Y=1)$ of a question, to which respondents evade giving a truthful answer, would be often small. In addition, an indirect questioning technique is used in order to ensure protection of privacy. The respondents feel that their privacy is secured when many non-key items are included (Hubbard *et al.* 1989). The simulation studies show that in such situations, the cross-based method or double cross-based method is more efficient than the traditional stratified method.

The domain estimators obtained by the traditional stratified method are generally inconsistent with the estimator $\hat{\pi}$ as shown in (10). Poststratified estimator $\hat{\pi}_{ps}$ by the domain variable addressed is essential in order to ensure consistency. Alternatively, we have to divide the total sample into two subgroups so that the distributions of their domain variable match in advance. On the contrary, the domain estimators obtained by the cross-based and the double cross-based methods are consistent with $\hat{\pi}$ as shown in (21). However, it does not mean that the cross-based method automatically adjusts the two subgroups so that the sample distributions of the domain variable match between the two subgroups. For the cross-based method, post-stratification by the domain variables or other demographic variables is also admissible, but not indispensable.

Even when the double cross-based method is used, negative domain estimates are sometimes observed. It is

possible to avoid negative estimates by letting a negative estimate q_c of Q_c in (23) be zero. However, such an adjustment produces a positive bias in $p(Y=1|Z=z)$.

The data of the survey of the Japanese national character, which were used in the simulation experiments, are neither sensitive nor were they obtained via the item count technique. In the future, the performance of the proposed method should be assessed by applying it to data obtained via the item count technique.

Acknowledgements

The author is grateful to two anonymous reviewers and an assistant editor for their helpful comments on a previous version of this paper.

References

- Abul-El, Abdel-Latif, A., Greenberg, B.G. and Horvitz, D.G. (1967). A multiproportions RR model. *Journal of the American Statistical Association*, 62, 990-1008.
- Chaudhuri, A., and Mukerjee, R.M. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons, Inc.
- Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W. and Ezzati, T.M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In *Measurement Errors in Surveys* (Eds. P.P. Biemer, *et al.*), New York: John Wiley & Sons, Inc.
- Droitcour, J.A., Larson, E.M. and Scheuren, F.J. (2001). The three card method: Estimating sensitive survey items with permanent anonymity of response. *Proceedings of the Social Statistics Section of the American Statistical Association*. Alexandria, V.A.: American Statistical Association.
- Greenberg, B.G., Abul-El, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question RR model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.
- Hubbard, M.L., Casper, R.A. and Lessler, J.T. (1989). Respondent reactions to item count lists and randomized response. *Proceedings of the Survey Research Section of the American Statistical Association*. Washington, D.C.: American Statistical Association. 544-548.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Lessler, J.T., and O'Reilly J.M. (1997). Mode of interview and reporting sensitive issues: Design and implementation of audio computer-assisted self-interviewing. *NIDA Research Monograph*, 167, 366-382.
- Miller, J.D. (1985). The nominative technique: A new method of estimating heroin prevalence. *NIDA Research Monograph*, 57, 104-124.
- Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey: John Wiley & Sons, Inc.

- Sakamoto, Y., Tsuchiya, T., Nakamura, T., Maeda, T. and Fouse, D.B. (2000). *A Study of the Japanese National Character: The Tenth Nationwide Survey (1998)*. Tokyo: The Institute of Statistical Mathematics Research Report General Series 85.
- Särndal, C.-E., Swesson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Schuman, H., and Presser, S. (1981). *Questions & Answers in Attitude Surveys*. New York: Academic Press.
- Takahasi, K., and Sakasegawa, H. (1977). A randomized response technique without making use of any randomizing device. *Annals of the Institute of Statistical Mathematics*, 29, 1-8.
- U.S. General Accounting Office (1999). *Survey Methodology. An Innovative Technique for Estimating Sensitive Items*. Washington D.C.: General Accounting Office.
- Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

Editing Systematic Unity Measure Errors Through Mixture Modelling

Marco Di Zio, Ugo Guarnera and Orietta Luzi¹

Abstract

In Official Statistics, data editing process plays an important role in terms of timeliness, data accuracy, and survey costs. Techniques introduced to identify and eliminate errors from data are essentially required to consider all of these aspects simultaneously. Among others, a frequent and pervasive systematic error appearing in surveys collecting numerical data, is the unity measure error. It highly affects timeliness, data accuracy and costs of the editing and imputation phase. In this paper we propose a probabilistic formalisation of the problem based on finite mixture models. This setting allows us to deal with the problem in a multivariate context, and provides also a number of useful diagnostics for prioritising cases to be more deeply investigated through a clerical review. Prioritising units is important in order to increase data accuracy while avoiding waste of time due to the follow up of non-really critical units.

Key Words: Editing; Random error; Systematic error; Selective editing; Model-based cluster analysis.

1. Introduction

Elements determining the quality of an Editing and Imputation (E&I) process are various and have been widely discussed in literature (Granquist 1995). We deal with a particular non-sampling error that highly affects two main competing quality dimensions: timeliness and data accuracy. As far as accuracy is concerned, we adopt the definition suggested in the Encyclopedia of Statistical Sciences, (1999): “accuracy concerns the agreement between statistics and target characteristics”. A number of factors can cause inaccuracy along the overall statistical survey process. Inaccuracy can be reduced during the E&I phase, which can be viewed as an “accuracy improvement tool by which erroneous or highly suspect data are found, and if necessary corrected (imputed)” (Federal Committee on Statistical Methodology 1990).

Due to the complexity of investigated phenomena and the existence of several types of non-sampling errors the E&I phase can be a very complex and time consuming task (Granquist 1996). In the specialised literature a common error classification leads to define two different error typologies: *systematic error* and *random error*. The former relates to errors which go in the same direction and lead to a bias in statistics, while the latter refers to errors which spread randomly around zero and affect the variance of estimates (Encyclopedia of Statistical Sciences 1999). Understanding nature of errors is not only useful in order to identify their source and to assess their effects on estimates, but also to adopt the most appropriate methodology to deal with them (Di Zio and Luzi 2002). While the Fellegi–Holt approach (Fellegi and Holt 1976) is a well-established paradigm to deal with random errors, systematic errors are generally treated by means of ad hoc solutions (see for

instance Euredit 2003, Vol. 1, Chapter 5). Systematic errors are generally treated before dealing with random errors, particularly when the latter are tackled through automatic software, like for instance the Generalised Editing and Imputation System (GEIS) (Kovar, Mac Millan and Whitridge 1988) and more recently De Waal (2003).

In the family of systematic errors, one that has a high impact on final estimates and that frequently affects data in statistical surveys measuring quantitative characteristics (e.g., business surveys) is the *unity measure error times a constant factor* (e.g., 100 or 1,000). This error is due to the erroneous choice, by some respondents, of the unity measure in reporting the amount of some questionnaire items.

As real examples of surveys affected by this type of error, we selected two ISTAT investigations: the 1997 Italian Labour Cost Survey (LCS) and the 1999 Italian Water Survey System (WSS).

The LCS is a periodic sample survey that collects information on employment, worked hours, wages and salaries and labour cost on about 12,000 enterprises with more than 10 employees. In Figure 1 the logarithm of Labour Cost (LCOST), Number of Employees (LEMPLOY), Worked Hours (LWORKEDH) are represented in a scatter plot matrix. Note that the employment variable at this editing stage is error free because of a preliminary check with respect to information from business registers (Cirianni, Di Zio, Luzi and Seeber 2000). The analysis of Figure 1 shows that Labour Cost is affected by two types of unity measure error (i.e., 1 million and 1,000 factor), while Worked Hours exhibits only the 1,000 factor error. These errors cause the different clusters in Figure 1. Note that the clusters in the low left corners of each scatter plot represent non-erroneous data.

1. Marco Di Zio, Ugo Guarnera and Orietta Luzi, Italian National Statistical Institute, Via Cesare Balbo 16, 00184 Roma, Italy.



Figure 1. Multiple scatter plot between total labour cost, employees, worked hours (logarithmic scale).

The WSS example will be described in detail in subsection 4.2 where an application of the method proposed in this paper for identifying and treating the unity measure error will be presented.

For the unity measure error, the critical point is the localisation of items in error rather than their treatment. In fact, once an item is classified as erroneous, the optimal treatment is uniquely determined and consists in a deterministic action recovering the original value through an inverse action (e.g., division by 1,000) neutralising the error effect.

The unity measure error is generally tackled through *ad hoc* procedures using essentially graphical representations of marginal or bivariate distributions, and *ratio edits*. A ratio edit is a rule stating that the value of a ratio between two variables must lie within a predefined interval. The interval bounds are generally determined through a priori knowledge or via exploratory data analysis, possibly using reliable auxiliary information. For this type of error, ratio edits are effective when one of the two variables is error free. Furthermore ratio edits allow taking into account only bivariate relationships between variables and even using interactive graphical inspection (e.g., scatter plot matrix), no more than a pairwise analysis can be performed, disregarding more complex interactions between variables. Finally, we notice that adopting pairwise analyses implies that variables are to be treated in a pre-defined hierarchy, thus increasing the complexity of the error localisation procedure.

With traditional approaches, the error localisation problem is not only complex, but also time and cost consuming. Time and cost are mainly affected by: 1) the complexity of designing and implementing automatic deterministic *ad hoc* procedures, and 2) the resources spent in manually editing

observations having low probabilities of being in error and/or low impact on target estimates (*over-editing*).

In this paper we propose a probabilistic formalisation of the problem through finite mixture models (McLachlan and Basford 1988; McLachlan and Peel 2000).

This modelling can provide a principled statistical approach, allowing an estimate of the conditional probability that an observation be affected by unity measure error. The advantage of the proposed approach is that it represents a general method allowing a multivariate data analysis, and providing elements that can be used to optimise the balance between the automatic and interactive components of the editing procedure, i.e., between time and accuracy (Granquist and Kovar 1997).

This work is organised as follows. In section 2 the proposed model is introduced together with the EM algorithm for the estimates of the model parameters. In section 3 diagnostics for selective editing are described. In section 4 the results of the application of the proposed method to both simulated and real data are illustrated. Finally, in section 5 concluding remarks and future research are outlined.

2. The Model

It is hard to give a comprehensive formalisation of random and systematic errors. In this context, we provide a definition that, though not exhaustive, includes many common situations. Let X^* be the vector of the survey target variables, and (μ, Σ^*) the corresponding mean vector and covariance matrix. Let us suppose that the measurement process is affected by a random error mechanism R having impact on the covariance structure of X^* but leaving the mean vector unchanged, and consequently let X be the corresponding “contaminated” variable, with $E(X) = E(X^*) = \mu$, $\text{Var}(X) = \Sigma$. Also, we assume that X can in turn be affected by a systematic error mechanism S acting only on its expected value: $\mu \xrightarrow{S} \phi(\mu)$ for some function ϕ (e.g., if an additive error mechanism is assumed, $\phi(\mu) = \mu + \text{constant}$). As a consequence of the two error mechanisms, assumed to be independent of one another, observed data can be described by a random vector Y whose distribution, conditional on X , depends only on the systematic error mechanism. Our approach to the treatment of systematic errors consists of building up a model for Y focusing only on the detection of systematic errors, thus aiming at recovering the randomly contaminated data represented by the random vector X . This is the approach generally adopted in editing procedures, where systematic errors and random errors are dealt with separately and hierarchically.

The previous definition of systematic error includes unity measure error, once data have been transformed in logarithmic scale. In fact, unity measure error generally acts multiplying variables by a constant factor. Hence data in error appear in log-scale as translated by a vector of constants, that depends on which items are in error ("error pattern"), while the covariance structure is the same for each error pattern. Moreover, as matter of fact, in business surveys variables are frequently considered log-normal. Thus in logarithmic scale the Gaussian setting can be adopted.

Following the formalisation so far introduced, our goal becomes to assign each single observation to a specific "error pattern", that corresponds to localise items in error. If we interpret each single error pattern as a "cluster", the error localisation problem is transformed in a cluster analysis problem, and we can exploit experiences from the model-based cluster analysis theory (Fraley and Raftery 2002).

More in detail, let us suppose we have n independent observations $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iq})$, $i=1, \dots, n$, corresponding to the q -dimensional vectors $\mathbf{X}_i = (X_{i1}, \dots, X_{iq})$ with p.d.f. $f(x_1, \dots, x_q; \theta)$, such that $E(X_1, \dots, X_q) = (\mu_1, \dots, \mu_q) = \boldsymbol{\mu}$, and $\text{Var}(X_1, \dots, X_q) = \Sigma$.

Based on the assumption that systematic errors affect the random vector \mathbf{X} only by transforming its expected value $\boldsymbol{\mu}$ into $\boldsymbol{\varphi}_g(\boldsymbol{\mu})$, where $\boldsymbol{\varphi}_g(\cdot): \mathbb{R}^q \rightarrow \mathbb{R}^q$, for $g=1, \dots, h$, are a set of known functions, the functions $\boldsymbol{\varphi}_g$ characterise univocally h distinct clusters (error patterns), differing each other only on the location parameter. For instance, if the systematic error possibly affects all the variables X_s for $s=1, \dots, q$, in the same manner by transforming their expected values μ_s according to $\mu_s \rightarrow \mu_s + C$, where C is a known constant, the number of clusters will be $h=2^q$, i.e., the number of different combinations of error occurrence on the q variables (including the case of no error). In this case, each function $\boldsymbol{\varphi}_g$ and each corresponding cluster, is associated with one of the 2^q possible sub-sets of variables affected by the error; e.g., the group G characterised by the mean vector $\boldsymbol{\mu}_G = (\mu_1, \mu_2 + C, \mu_3, \mu_4, \dots, \mu_q)$, is a cluster of units with error affecting only the variable X_2 . We remark that we assume a common covariance matrix because we make the hypothesis that the possible random error acts in the same way on all the data.

For the error localisation purpose we follow a model-based approach based on finite mixture models, where each mixture component G_g , $g=1, \dots, h$, represents a single error pattern. Formally, we assume that $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iq})$, for $i=1, \dots, n$, are iid w.r.t. $\sum_{t=1}^h \pi_t f_t(\cdot; \theta_t)$, where $\sum_t \pi_t = 1$ and $\pi_t \geq 0$. The mixing parameters π_t represent the probability that an observation belongs to the t^{th} mixture component.

In order to classify an observation \mathbf{y}_i in one of the h groups, we compute the posterior probability $\tau_g(\mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) = \text{pr}(i^{\text{th}} \text{ observation} \in G_g | \mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi})$, that is

$$\tau_g(\mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) = \pi_g f_g(\mathbf{y}_i; \boldsymbol{\theta}_g) / \sum_{t=1}^h \pi_t f_t(\mathbf{y}_i; \boldsymbol{\theta}_t) \quad g=1, \dots, h. \quad (1)$$

The i^{th} observation is assigned to the cluster G_g , if

$$\tau_t(\mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) > \tau_g(\mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) \quad g=1, \dots, h; g \neq t.$$

The previous allocation rule is the optimal solution for the classification problem, in the sense that it minimises the overall error rate (Anderson 1984, Chapter 6).

Since, in place of the parameters $(\boldsymbol{\theta}, \boldsymbol{\pi})$, generally unknown, we use the maximum likelihood estimates $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\pi}})$, the classification rule becomes:

$$\tau_t(\mathbf{y}_i; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\pi}}) > \tau_g(\mathbf{y}_i; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\pi}}) \quad g=1, \dots, h; g \neq t. \quad (2)$$

We assume that the $f_t(\mathbf{y}; \boldsymbol{\theta}_t)$ is a multivariate normal density $MN(\boldsymbol{\mu}_t, \Sigma)$ and that each function $\boldsymbol{\varphi}_g(\cdot)$ acts on the mean vector $\boldsymbol{\mu}$ as a translation: $\boldsymbol{\varphi}_g(\boldsymbol{\mu}) = \boldsymbol{\mu} + \mathbf{C}_g$, where \mathbf{C}_g represents the translation vector for the mean of the g^{th} cluster, and it is supposed to be known. This setting, as already noticed, is suitable for dealing with unity measure error. In order to compute the likelihood estimates, we use the EM algorithm as suggested in McLachlan and Basford (1988). Nevertheless, an additional effort is necessary to adapt the algorithm to our particular situation, where the mean vectors of the mixture components are linked by a known functional relationship. Thus, while in the non-constrained case (McLachlan and Basford 1988) a different mean vector has to be estimated for each mixture component, in our constrained situation only one mean vector needs to be estimated. The resulting modified EM algorithm consists of defining some initial guess for the parameters to be estimated $\hat{\boldsymbol{\mu}}_g^{(0)}$ for $g=1, \dots, h$, $(\hat{\boldsymbol{\mu}}^{(0)}, \hat{\Sigma}^{(0)})$ and applying until convergence the following recursive scheme:

- i) compute the posterior probabilities $\tau_{gi}^{(k)} = \tau_g^{(k)}(\mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi})$ under the current estimates $\hat{\boldsymbol{\pi}}^{(k)}$, $\hat{\boldsymbol{\mu}}^{(k)}$, $\hat{\Sigma}^{(k)}$ (k is the index referring to the k^{th} cycle)

$$\begin{aligned} & \hat{\tau}_{gi}^{(k)} \\ &= \frac{\hat{\pi}_g^{(k)} \exp \left\{ -\frac{1}{2} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_g^{(k)})' \left(\hat{\Sigma}^{(k)} \right)^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_g^{(k)}) \right\}}{\sum_{t=1}^h \hat{\pi}_t^{(k)} \exp \left\{ -\frac{1}{2} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_t^{(k)})' \left(\hat{\Sigma}^{(k)} \right)^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_t^{(k)}) \right\}} \end{aligned}$$

ii) calculate the new estimates by the following recursive equations:

$$\begin{aligned}\hat{\pi}_g^{(k+1)} &= \sum_{i=1}^n \hat{\tau}_{gi}^{(k)} / n \\ \hat{\mu}^{(k+1)} &= \sum_{g=1}^h \sum_{i=1}^n \hat{\tau}_{gi}^{(k)} y_i / n - \sum_{g=1}^h C_g \hat{\pi}_g^{(k+1)} \\ \hat{\Sigma}^{(k+1)} &= \sum_{g=1}^h \sum_{i=1}^n \hat{\tau}_{gi}^{(k)} (y_i - \mu_g^{(k+1)})(y_i - \mu_g^{(k+1)})' / n \hat{\pi}_g^{(k+1)}.\end{aligned}$$

We remark that $\hat{\mu}_g^{(k)}$ stands for $\hat{\mu}^{(k)} + C_g$.

In practical applications, it turns out that a crucial role is played by the choice of starting points, as usual in the EM algorithms (see Biernacki, Celeux and Govaert 2003). To overcome this problem, we use an initialisation strategy, following Biernacki *et al.* (2003), consisting of several short runs in terms of number of iterations, of the algorithm from random initialisations followed by a long run of EM from the solution maximising the observed log-likelihood.

It is worth to mention that, due to the location constraints, the parameters to be estimated are sensibly fewer than those in a usual mixture problem. Actually the higher is the number of variables analysed the bigger is this difference; for instance in the case of three variables and 8 clusters we need to estimate 16 parameters instead of 37. This aspect is particularly important when we deal with small samples. Moreover, constraints on cluster locations make easier to identify “rare clusters”. In fact, being the relative distances between mean vectors fixed, the estimation problem reduces to estimate the location of the convex polyhedron whose vertices are the cluster centroids. In other words, since the location of one centroid univocally determines the positions of all the others, small cluster parameters are more easily estimated than if they were not constrained.

Since the introduced modelling is based on the assumption that observations are normally distributed, model validation is an issue to take into account. The problem of assessing normality in mixture models is well described in McLachlan and Basford (1988). It is essentially based on the quantities \hat{a}_{gi} described in the following. Let y_{gi} for $i=1, \dots, \hat{m}_g$ be the observations assigned to the g^{th} cluster for $g=1, \dots, h$, according to the estimated model. Let \hat{p}_{gi} be the value calculated using the estimated parameters, following the formula:

$$\hat{p}_{gi} = \frac{(v\hat{m}_g / q) D(y_{gi}, \hat{\mu}_g; \hat{\Sigma})}{(v+q)(\hat{m}_g - 1) - \hat{m}_g D(y_{gi}, \hat{\mu}_g; \hat{\Sigma})}, \quad (3)$$

where $D(\cdot, \cdot; M)$ is the Mahalanobis squared distance based on the metric M , and $v = n - h - q$. We define \hat{a}_{gi}

as the area to the right of the \hat{p}_{gi} value under the $F_{q,v}$ distribution (for details see McLachlan and Basford 1988, Chapter 2).

Under the normality assumption, \hat{a}_{gi} for $i=1, \dots, \hat{m}_g$ is approximately uniformly distributed on $(0,1)$. Hawkins (1981) suggests using the Anderson–Darling statistic for assessing the uniform distribution of \hat{a}_{gi} . The \hat{a}_{gi} are also useful to detect outliers, *i.e.*, atypical observations with respect to the model. In McLachlan and Basford (1988) the lower is \hat{a}_{gi} the higher is the probability of y_{gi} of being atypical, thus all observations with $\hat{a}_{gi} < \alpha$, where α is a specified threshold, can be considered as atypical. Suggested threshold levels range from $\alpha=0.05$ to $\alpha=0.005$, depending on which outlying observations (more or less extreme values) are to be selected.

3. Diagnostics for Selective Editing

Once the parameters of the mixture have been estimated, we are able to classify data into the different clusters, *i.e.*, for each observation we can assess whether it is in error or not, and which variables are in error. However, different types of critical observations can be identified after the modelling phase: units classified in a cluster, but having a non-negligible probability of belonging to another cluster, and observations that are outliers with respect to the model.

In order to increase data accuracy it would be useful to make a double check on critical observations (through either a clerical review or, in the most difficult cases, a follow-up). On the other hand, in order to reduce possible over-editing and editing costs, the manual review and/or follow up should be concentrated on the most critical observations. The proposed mixture model directly provides diagnostics that can be used to this aim.

A first type of critical units is represented by possibly misclassified observations. In order to measure the degree of belief in the class assigned to an observation y_i we can consider the corresponding probability resulting from (2). Observations, for which this probability is not very close to one, have a non-negligible probability to belong to another cluster. These observations are those in the region where the mixture components overlap each other.

In addition to the previous type of critical units, there are other observations that are far from all the clusters (all the mixture components), *i.e.*, outliers with respect to the model. Also these observations represent critical situations. In order to identify this kind of outlier we refer to the quantities \hat{a}_{ij} described in the previous section.

Classification probability and atypicality index \hat{a}_{gi} should be used, according to a selective/significance editing approach (Latouche and Berthelot 1992; Lawrence and McKenzie 2000), to build up appropriate score functions to

prioritise critical units. An example of how to use these diagnostics to this aim is given in subsection 4.2.

4. Illustrative Examples

In this section some experiments carried out in order to investigate the peculiarities of the proposed method are presented. Firstly, through a simulation study, we analyse the performance of the proposed model when applied to data that depart from normality. Secondly, through an application on real data, we describe how this approach can be applied in Official Statistics.

All the experiments are performed using the R environment for statistical computing (<http://www.r-project.org/>).

4.1 Simulated Example: Departure from Normality

In this experiment we describe the results obtained by applying the mixture approach to the three different populations depicted in the first line of Figure 2. The first distribution is a bivariate normal (MN), hence it represents the case when the model is correctly specified. The second one corresponds to a bivariate *t* distribution (MT), *i.e.*, it mimes the situation when the departure from normality is essentially in having heavier tails. The last one is a bivariate skew-*t* distribution (ST) (Azzalini and Capitanio 2003, Azzalini, Dal Cappello and Kotz 2003), and it represents a

population distributed according to an asymmetric distribution with heavy tails.

From these distributions we build a four components mixture model by adding to each unit one of the four translation vectors $C_1 = (0, 0)$, $C_2 = (0, \log(1,000))$, $C_3 = (\log(1,000), 0)$, $C_4 = (\log(1,000), \log(1,000))$ with probabilities $\pi_1 = 0.5$, $\pi_2 = 0.1$, $\pi_3 = 0.1$, and $\pi_4 = 0.3$ respectively. These parameters represent the mixing proportions of the mixture model and refer respectively to the probabilities of no translation in the variables, translation in only one of the two variables, and translation in both variables. From each mixture, we draw 100 samples of 1,000 observations. In the second line of Figure 2, we report one of these samples (MN-Mixt, MT-Mixt, ST-Mixt), corresponding to the three different populations MN, MT, ST respectively.

For each sample, we compute the number of correct classifications obtained by using the mixture approach described in section 2. The mean number of correct classifications over the 100 samples is reported in Table 1.

As it can be seen in Table 1, the frequency of correct classifications decreases with the departure from normality. However it seems acceptable also in the critical case ST, where the population is characterised by both asymmetry and heavy tails.

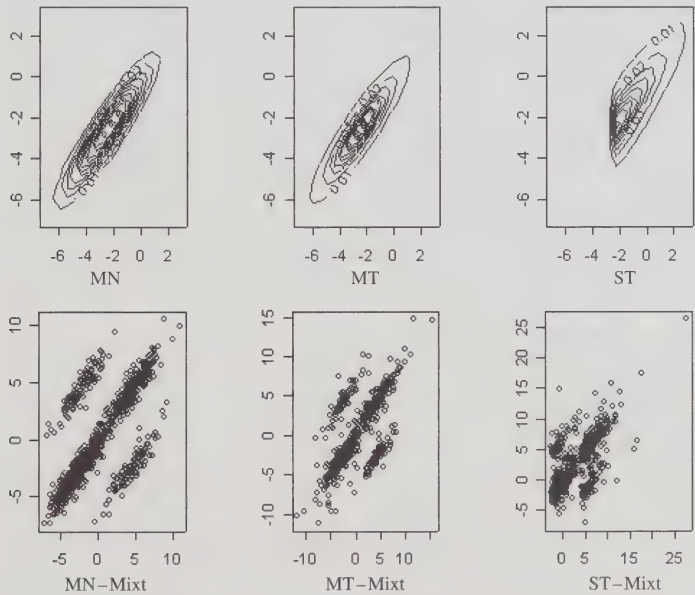


Figure 2. Contour plots of the three bivariate distributions multinormal (MN), *t*-student (MT), skew-*t* (ST), and scatter plot of the corresponding mixtures MN-Mixt, MT-Mixt, ST-Mixt.

Table 1
Frequency of Correct Classifications

	MN	MT	ST
% correctly classified	98.5	97.5	95.6

As discussed in section 3, the mixture approach provides elements (such as the degree of atypicality and the classification probability) that can be used in order to prioritise units to be clerically reviewed. Therefore, an overall assessment of the procedure should consider also the results obtained through a selective editing approach based on these model diagnostics.

In order to analyse the characteristics of atypicality index and classification probability, we examine a single sample of 1,000 observations drawn from the three populations so far introduced. In Figure 3, the three samples MN-Mixt(a), MT-Mixt(a), ST-Mixt(a) are represented, furthermore the misclassified units are depicted with a cross in the same graph. The number of misclassified units is 19 for MN-Mixt, 20 for MT-Mixt, and 36 for ST-Mixt.

On this sample, we focus on the impact of different threshold levels both for atypicality (α) and classification probability (β). For each threshold, we report in Table 2 and Table 3 the number of units below that threshold, *i.e.*, the number of critical observations (*N. Atyp*, *N. Pr. Class*),

and among them the number of misclassified units (*Atyp - Misclas*, *Pr. Class - Misclas*).

As far as atypicality is concerned, we note that when the model is correctly specified, the importance of the atypicality index in recovering misclassified units is negligible, while the classification probabilities are more effective. On the other hand the degree of atypicality is important when the model departs from normality. It is clear that the number of observations selected for a given combination of thresholds α and β is not equal to the sum of the frequencies obtained in Table 2 and Table 3. Thus, in order to evaluate the joint impact of these two indices we choose the two following thresholds $\alpha=0.005$ and $\beta=0.975$. We report in Figure 3 (second line) the units selected only for the atypicality value (squares), only for the classification probability (triangles), and for both of them (crosses). From these figures we see how the impact of atypicality is mainly on outliers identification while the classification probability works on the overlapping regions. In Table 4 the number of selected units and, out of them the number of misclassified units are shown.

We note that for population MN-Mixt, apart one observation, all the misclassified units are selected. For MT-Mixt, we are able to select 14 out of the 20 misclassified units, and in the most critical sample ST-Mixt we select 24 out of the 36 misclassified units.

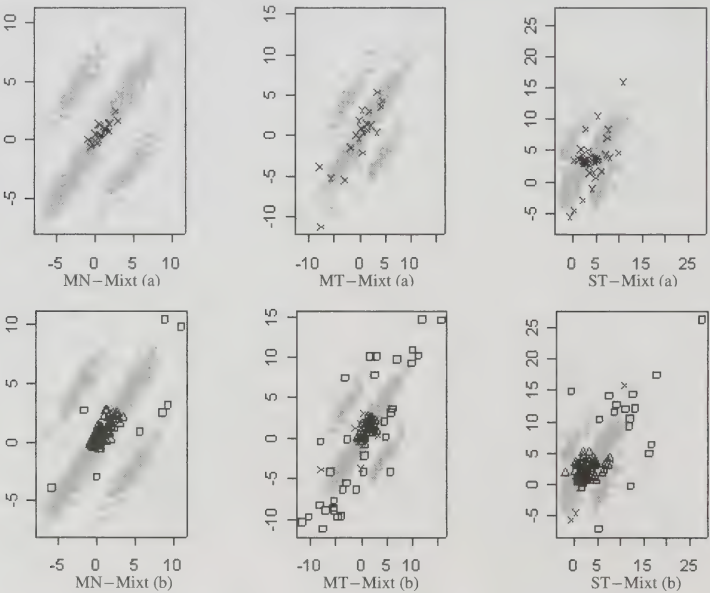


Figure 3. Misclassified units (crosses) in MN-Mixt(a), MT-Mixt(a), ST-Mixt(a). Critical units for atypicality (square), for classification probability (triangle), and for both of them (cross), in MN-Mixt(b), MT-Mixt(b), ST-Mixt(b).

Table 2

Number of Critical Observations and Misclassified Units with Respect to Three Different Thresholds for Atypicality

α	MN-Mixt		MT-Mixt		ST-Mixt	
	<i>N. Atyp</i>	<i>Atyp – Misclas</i>	<i>N. Atyp</i>	<i>Atyp – Misclas</i>	<i>N. Atyp</i>	<i>Atyp – Misclas</i>
0.05	50	1	84	9	68	14
0.01	15	0	50	7	33	8
0.005	8	0	39	7	20	5
0.001	4	0	25	4	14	2

Table 3

Number of Critical Observations and Misclassified Units with Respect to Three Different Thresholds for Classification Probability

β	MN-Mixt		MT-Mixt		ST-Mixt	
	<i>N. Pr. Class</i>	<i>Pr. Class – Misclas</i>	<i>N. Pr. Class</i>	<i>Pr. Class – Misclas</i>	<i>N. Pr. Class</i>	<i>Pr. Class – Misclas</i>
0.99	119	19	63	12	182	26
0.975	76	18	46	11	82	26
0.95	55	14	35	9	66	21

Table 4

Number of Critical Observations and Misclassified Units with Respect to Atypicality and Classification Probability

<i>Thresholds</i>	MN-Mixt		MT-Mixt		ST-Mixt	
	<i>N.Crit. Units</i>	<i>N. Misclas</i>	<i>N.Crit. Units</i>	<i>N. Misclas</i>	<i>N.Crit. Units</i>	<i>N. Misclas</i>
$\alpha = 0.005, \beta = 0.975$	84	18	79	14	98	24

4.2 An Application to Real Data: The 1999 Italian Water Survey System

In this section we describe an application of the mixture model approach to real survey data. The data are taken from the 1999 Italian *Water Survey System* (WSS). The WSS is a census that collects information on water abstraction, supply and usage for the 8,100 Italian municipalities. We restrict our analysis to the municipalities belonging to one of the data domains defined by altimetry (2,041 observations) and to the main variables *Total Invoiced Water* (TI) and *Total Supplied Water* (TS). Both these variables refer to water volumes and the respondents are requested to provide them in thousands of cubic meters. The scatter plot on log-scale of per capita water invoiced (WI) versus per capita water supplied (WS) (Figure 4) shows the presence of four clusters corresponding to unity measure error in one, both, or none of the target variables. This is probably due to the misunderstanding of some respondents that expressed water volumes in litres or in cubic meters rather than thousands of cubic meters, as requested. As expected, the two most populated clusters are those corresponding to non-erroneous units and to units where both variables are in error. Nevertheless, we can note the presence of two rare clusters corresponding to observations where the unity measure error affects only TI or only TS respectively.

In Table 5 a label is assigned to each group associated with a specific error pattern. For the sake of simplicity we introduce two flags E_{TS} and E_{TI} assuming value 1 or 0,

depending on whether the corresponding variables are affected by the unity measure error or not, respectively.

In order to identify and correct the unity measure error we apply the procedure described in sections 2 and 3. We classify each observation according to a specific error pattern, *i.e.*, we assign each unit to one of the clusters G_t , for $t = 1, \dots, 4$. The results are reported in Table 6.

For each unit the atypicality index is also calculated and the threshold $\alpha = 0.005$ is chosen in order to flag atypical units. According to this threshold, 71 observations are selected as atypical, marked by “crosses” in Figure 7. Once the values \hat{a}_{gi} are computed according to Formula (3), a test assessing the normality assumption can be performed. Actually, following McLachlan and Basford (1988, Chapter 2), the Anderson–Darling test on the uniformity of \hat{a}_{gi} on each single estimated cluster is performed. The p -values are below 0.001 for the two largest clusters. Since the test is based on asymptotical approximations, we do not take into account the results on the other two rare populations. In Figure 5 we report the empirical sample quantiles versus the normal quantiles of the variables $\log(WI)$ and $\log(WS)$, focusing only on the subset of data classified as non-erroneous. We notice that departure from normality is mainly due to heavy tails. Based on the results obtained in section 4.1, where the method performed satisfactorily also in non-gaussian setting, we are confident about the good performance of the mixture approach on the survey data. This expected behaviour is confirmed by the application results showed in the following.

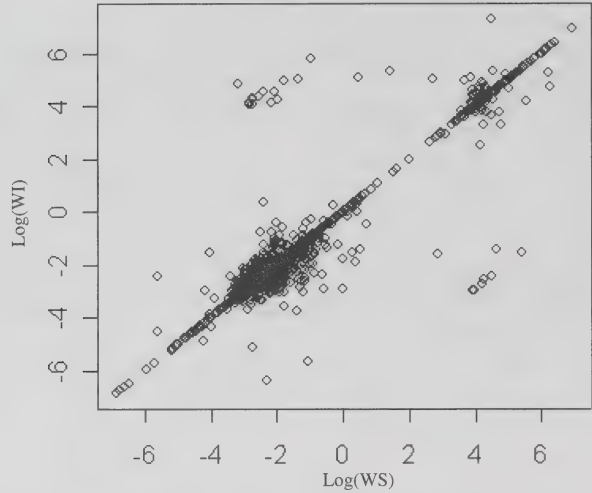


Figure 4. Scatter plot of log(WS) and log(WI).

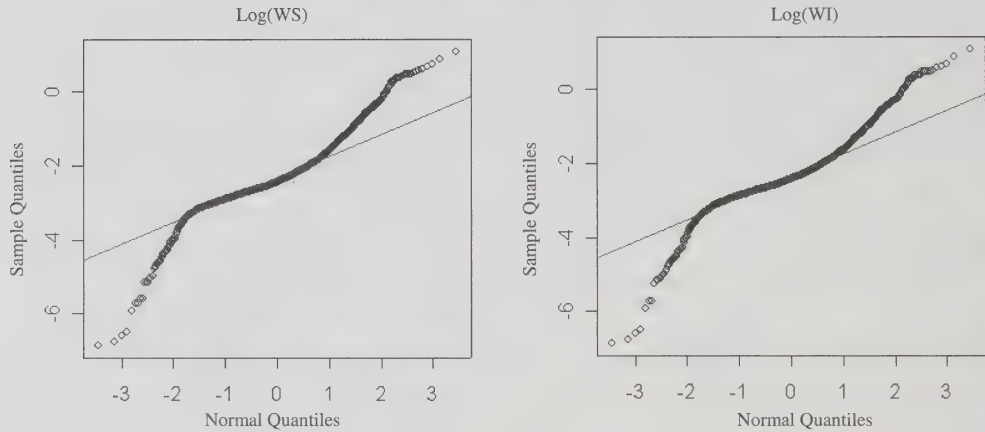


Figure 5. Normal qq-plot of log(WS) and log(WI).

Table 5

Error Patterns and Error Labels

Error pattern	$E_{TS} = 0$	$E_{TS} = 0$	$E_{TS} = 1$	$E_{TS} = 1$
	$E_{TI} = 0$	$E_{TI} = 1$	$E_{TI} = 0$	$E_{TI} = 1$
Cluster label	G1	G2	G3	G4

Table 6

Number of Units Assigned to Each Cluster

Cluster label	G1	G2	G3	G4
N. of units	1,800	16	10	215
%	88.2	0.8	0.5	10.5

In the remaining part of this section, it is shown how the posterior probabilities can be used to prioritise units to be reviewed which are likely to provide the greatest editing benefit, taking into account the potential impact of the clerical editing on the estimates. To this aim, note that a wrong classification of an observation causes that the final values of at least one variable differ from the corresponding true values by a multiplicative factor. These discrepancies can seriously affect the accuracy of the estimates leading to a strong bias. In order to select the potentially erroneous units that most likely have a strong impact on the target estimates, we follow the *selective editing approach*. Let X_1 , X_2 denote the variables *TS*, *TI* respectively. For each unit u_i , $i = 1, \dots, n$, and for each variable X_j , $j = 1, 2$, let us define:

X_{ij} : data free of systematic error;

Y_{ij} : observed data;

\tilde{X}_{ij} : data after the treatment of systematic error based on the classification through mixture model (*i.e.*, $\tilde{X}_{ij} = Y_{ij}$ or $\tilde{X}_{ij} = Y_{ij}/1,000$ depending on the cluster the unit u_i is assigned to).

Let us suppose that the target estimates refer to population totals $T(X_j) = \sum_i X_{ij}$. Further, denote by $E_\xi(\cdot)$ the expectation over the distribution of the random variable X_j conditional on the observed data Y_{ij} and the data after correction \tilde{X}_{ij} . Then, from the inequality $|\sum_i E_\xi(X_{ij} - \tilde{X}_{ij})| \leq \sum_i E_\xi |X_{ij} - \tilde{X}_{ij}|$ it follows that the quantity on the right hand side can be viewed as an upper bound for the expected bias of the total estimate for the variable X_j based on the corrected values \tilde{X}_{ij} . The last consideration suggests a method for selecting the most "influential" units with respect to the estimate $T(X_j)$: in order to guarantee the requested level of accuracy and to minimise costs due to manual check, we define a local score function $S_{ij} = (E_\xi |X_{ij} - \tilde{X}_{ij}|) / \hat{T}(X_j)$, where $\hat{T}(X_j)$ is a reference estimate for $T(X_j)$, for instance the estimate from a previous survey, or a robust estimate. In our case, in order to robustify the preliminary estimate we first exclude from the data the atypical observations, then compute the mean value on this subset, and then multiply it by the total number of units.

The local score S_{ij} measures the impact of the potential unity measure error associated to the unit u_i on the target estimate $T(X_j)$. Then, units can be sorted by their score S_{ij} and, starting from the highest values, the first units can be selected until the sum of the remaining S_{ij} values is lower than a predefined threshold.

If both the variables TS and TI are considered simultaneously, a global score S_i , for $i=1, \dots, n$, can be obtained by suitably combining the local score functions S_{ij} , $j=1, 2$. Possible choices are $S_i = (S_{i1} + S_{i2})/2$, or $S_i = \max_{j=1, 2} S_{ij}$. The latter function, for instance, ensures that the impact of the potential unity measure error associated with u_i on each estimate is not greater than S_i .

In order to compute the scores S_{ij} the conditional expected value $E_\xi |X_{ij} - \tilde{X}_{ij}|$ is to be estimated for each unit u_i , $i=1, \dots, n$, and for each variable X_j for $j=1, 2$. This can be easily done through the posterior probabilities. For instance, suppose that the unit u_i has been assigned to the cluster G_2 . This means that, for this unit, the observed value of TS (Y_{i1}) has been considered correct, while the observed value of TI (Y_{i2}) has been flagged as affected by unity measure error (*i.e.*, multiplied by 1,000). The correction consists of dividing by 1,000 the observed value

of TI, *i.e.* ($\tilde{X}_{i1} = Y_{i1}$, $\tilde{X}_{i2} = Y_{i2}/1,000$). The conditional expected value $E_\xi |X_{ij} - \tilde{X}_{ij}|$ can be computed as follows:

$$\begin{aligned} E_\xi |X_{i1} - \tilde{X}_{i1}| &= |Y_{i1} - Y_{i1}| \Pr(u_i \in G_1 \cup G_2) \\ &\quad + \left| \frac{Y_{i1}}{1,000} - Y_{i1} \right| \Pr(u_i \in G_3 \cup G_4) \\ &= \frac{999}{1,000} Y_{i1} (\hat{\tau}_{3i} + \hat{\tau}_{4i}) \\ E_\xi |X_{i2} - \tilde{X}_{i2}| &= \left| \frac{Y_{i2}}{1,000} - \frac{Y_{i2}}{1,000} \right| \Pr(u_i \in G_2 \cup G_4) \\ &\quad + \left| Y_{i2} - \frac{Y_{i2}}{1,000} \right| \Pr(u_i \in G_1 \cup G_3) \\ &= \frac{999}{1,000} Y_{i2} (\hat{\tau}_{1i} + \hat{\tau}_{3i}), \end{aligned}$$

where $\hat{\tau}_s$ is the estimated probability that unit u_i belongs to cluster G_s . In a similar manner the score functions can be calculated for all the units.

In practice, in our application we sort the units by their global score S_i , $\max_{j=1, 2} S_{ij}$ (ascending order). Then we exclude from clerical review all the first observations such that their cumulative sum of S_i is below δ , where δ is a specified tolerance level for the impact on the estimates due to errors remaining in data. In Figure 6 the behaviour of the cumulative sum of S_i , $S_{(i)} = \sum_{k \leq i} S_k$, is shown for the first most critical 10 observations. We remark that for the sake of clarity we have not reported all the observations because for most of them $S_{(i)}$ is close to zero causing an unreadable picture for their different magnitude. Note that a residual relative error less than $\delta = 0.001$ is expected by selecting only the first two units (drawn with crosses).

In Figure 7 all the units selected because of their atypicality (71) and/or the relative impact on estimates of their potential errors (2) are shown: crosses correspond to observations that are critical for atypicality, squares indicate the other two types of critical units.

A comparison with the results obtained by the official procedure is made. Out of the 1,968 units not selected for clerical review, 1,911 observations are error free or affected by unity measure error only. For all of them the classification of the mixture model is correct. Out of the remaining 57 units characterised by other error typologies, 45 are classified as non-affected by the unity measure error, while 12 as units with the 1,000 error in both the variables. This last misclassification can be explained by the presence of another systematic error (times 100, 10,000 factors) that is not taken into account in the model used for this example.

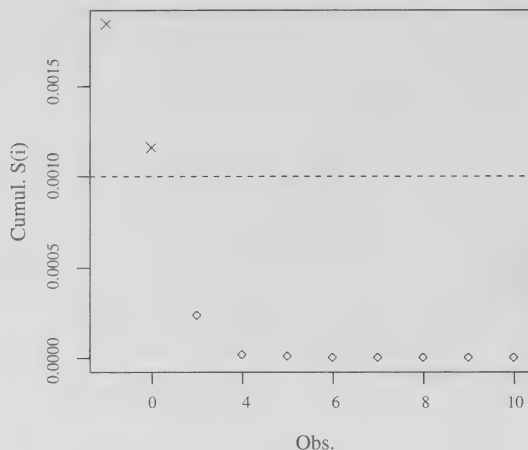


Figure 6. Plot of the cumulative score $S_{(i)}$ for the first most critical 10 observations.

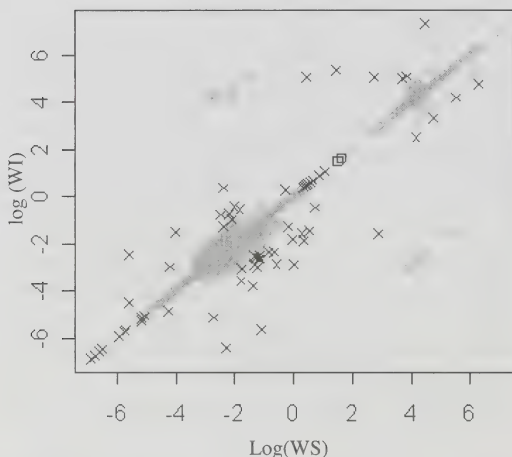


Figure 7. Scatter plot of $\log(\text{WS})$ vs $\log(\text{WI})$. Crosses indicate critical units for atypicality, squares mark critical units for the impact of their potential error.

A further comparison is about the estimate of the totals. Under the hypothesis that the values selected for a clerical review are correctly restored, the relative differences between the “true” total values according to the official procedure $T(X_j)$ and the model estimate $\hat{T}(X_j)$ as $B(X_j) = (|\hat{T}(X_j) - T(X_j)|) / T(X_j)$, for $j=1, 2$ are $B(X_1) = 0.005$ and $B(X_2) = 0.002$. These values are not directly comparable with the tolerance level $\delta = 0.001$, in fact this threshold relates only to impact of the remaining unity measure errors, while $B(X_j)$ is also affected by other

kind of errors. Thus, for a more direct comparison, we replace for these units the wrong values with the “true” ones obtaining $B(X_1) = B(X_2) = 0$. This particularly high performance of the model is justified by the low degree of overlapping of the clusters as clear in Figure 7.

5. Final Remarks and Further Research

In this paper we propose a finite mixture model to deal with a particular type of systematic error that frequently affects numerical continuous survey data: the unity measure error times a constant factor. The proposed approach has the advantages, with respect to the traditional ones, to formally state the problem in a multivariate context, to be easily implemented in generalised software, and to naturally provide useful diagnostics for prioritising doubtful units possibly containing influential errors. The latter characteristic is particularly important when the situation is critical, *i.e.*, when different error patterns overlap each other or in other words when unity measure errors are among plausible observations. In these circumstances a clerical review is needed. Hence, it is important to optimise the selection of critical observations in order to save time and costs. All these advantages are the natural consequence of the introduction of a model-based technique. On the other hand, it is clear that the use of a model-based approach implies problems related to model assumptions. However, based on the experiments illustrated in the paper, it seems that also in cases of departure from the normality assumption, the proposed technique performs satisfactorily. Nevertheless, it is worth to mention that for extreme departure from normality, *e.g.*, when the distribution is not unimodal, the method is expected to fail. This can happen in real situations when true data contain different clusters, for instance differences in men and women income might cause a bimodal distribution for the income itself. In some cases the problem could be overcome by stratifying data with respect to some explicative variables, *e.g.*, sex in the previous example. An alternative approach to this specific problem could be based on modelling each cluster in turn as a Gaussian mixture, thus obtaining a “mixture of mixture models” (McLachlan and Peel 2000; Di Zio, Guarnera and Rocci 2004).

Finally, a last concern is about the number of variables that can be treated simultaneously. Actually, the number of clusters and then the number of mixing parameters π , can have an exponential growth with respect to the number of variables, making the parameter estimation a critical task. However it is worthwhile noting that the number of parameters related to the mean vector and covariance matrix increases much slower, due to the constraints characterising our model.

Acknowledgements

We are grateful to the referees and the Associate Editor for their helpful comments.

References

- Anderson, T.W. (1984). *An introduction to Multivariate Statistical Analysis*. Second Edition. New York: John Wiley & Sons, Inc.
- Azzalini, A., and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew- t distribution. *Journal of the Royal Statistical Society (B)*, 65, 367-389.
- Azzalini, A., Dal Cappello, T. and Kotz, S. (2003). Log-skew-normal and log-skew- t distributions as models for family income data. *Journal of Income Distribution*, 11, 13-21.
- Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41, 561-575.
- Cirianni, A., Di Zio M., Luzzi O. and Seeber, A.C. (2000). The new integrated data editing procedure for the Italian Labour Cost survey: Measuring the effects on data of combined techniques. *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, 7-21.
- De Waal, T. (2003). Solving the error localization problem by means of vertex generation. *Survey Methodology*, 29, 1, 71-79.
- Di Zio, M., Guarnera, U. and Rocci, R. (2004). A mixture of mixture models to detect unity measure error. *Proceedings in Computational Statistics*, (Ed. Antoch Jaromir), 919-927, Physica Verlag, Prague, August 23-28.
- Di Zio, M., and Luzzi, O. (2002). Combining methodologies in a data editing procedure: an experiment on the survey of Balance Sheets of Agricultural Firms. *Italian Journal of Applied Statistics*, 14, 1, 59-80.
- Encyclopedia of Statistical Sciences (1999). New York: John Wiley & Sons, Inc. Update 3, 621-629.
- Euredit (2003). *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project*, 1, 2. Forthcoming. Now available at <http://www.cs.york.ac.uk/euredit/>
- Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18.
- Fellegi, I.P., and Holt, D. (1976). A systematic approach to edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.
- Fraley, C., and Raftery, A. (2002). Model-Based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.
- Granquist, L. (1995). Improving the traditional editing process. In *Business Survey Methods*, (Eds. B.G. Cox and D.A. Binder).
- Granquist, L. (1996). The new view on editing. *International Statistical Review*, 65, 3, 381-387.
- Granquist, L., and Kovar, J. (1997). Editing of survey data: How much is enough? In *Survey Measurement and Process Quality*, (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 415-435.
- Hawkins, D.M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics*, 23, 105-110.
- Kovar, J.G., Mac Millian, I.H. and Whitridge, P. (1988). Overview and strategy for the generalized edit and imputation system, (updated February 1991). Statistics Canada, Methodology Branch Working Paper, BSMD-88-007E/F.
- Latouche, M., and Berthelot, J.M. (1992). Use of a score function to prioritise and limit recontacts in business surveys. *Journal of Official Statistics*, 8, 389-400.
- Lawrence, D., and McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.
- McLachlan, G.J., and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.
- McLachlan G.J., and Peel D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

Using Matched Substitutes to Improve Imputations for Geographically Linked Databases

Wai Fung Chiu, Recai M. Yucel, Elaine Zanutto and Alan M. Zaslavsky¹

Abstract

When administrative records are geographically linked to census block groups, local-area characteristics from the census can be used as contextual variables, which may be useful supplements to variables that are not directly observable from the administrative records. Often databases contain records that have insufficient address information to permit geographical links with census block groups; the contextual variables for these records are therefore unobserved. We propose a new method that uses information from "matched cases" and multivariate regression models to create multiple imputations for the unobserved variables. Our method outperformed alternative methods in simulation evaluations using census data, and was applied to the dataset for a study on treatment patterns for colorectal cancer patients.

Key Words: Unit nonresponse; Multiple imputation; Contextual variables; Matched substitutes; Administrative records.

1. Introduction

In a study on treatment patterns for colorectal cancer patients, income and education are desired variables for constructing statistical models of relevant scientific interest. Unfortunately, individual measurements for these variables are not directly observable from the cancer registry databases that are compiled from hospital records, which like many administrative databases contain primarily information required for administrative purposes. Instead, mean values of these variables for small geographical areas (census block groups or tracts) including the subject's area of residence are used as regressors to estimate income and education effects. Analyses using such "contextual variables" are common in epidemiological and health services research (Krieger, Williams and Andmoss 1997), and often produce results broadly similar to those based on individual variables. If both individual and contextual variables were available, it might be possible to separate the effects of individual characteristics and contexts; in a purely contextual analysis, these effects are confounded. Nonetheless, associations between contextual socioeconomic characteristics and quality of care would suggest an equity problem, regardless of whether such associations primarily reflect individual or community-level relationships.

In the colorectal cancer treatment study, each contextual variable for a given patient record is assumed to be the variable's census group (or tract) mean value obtained by geographically linking the record's address to a census block group (or tract). A small but substantial percentage of

patient records (about 3.3% or 1,696 records) have insufficient address information to permit links with census block groups, hence making the corresponding contextual variables unobservable. Such records will be called *ungeocodable* records, while records that can be linked to census block groups will be referred to as *geocodable*. To generate multiple imputations for the unobserved contextual variables, we propose a strategy that uses information from more than one "matched case" to help build parametric/nonparametric imputation models. In particular, information from the matched cases accounts for small area effects in our imputation models, so that there is no need to explicitly model such effects.

Rubin and Zanutto (2001) use the term "matched substitute" instead of "matched case", and propose a parametric imputation model using only one matched substitute per record. The analyses resulted from their model were compared to those given by other analytic methods in an extensive simulation study, but was not applied to real data. We extend Rubin and Zanutto's method by (1) allowing use of information from more than one matched case per record and (2) using an empirical rather than a parametric distribution of residuals.

This research was motivated by our need for multiple imputations for the partially observed variables in the study of treatment patterns for colorectal cancer patients. Ayanian, Zaslavsky, Fuchs, Guadagnoli, Creech, Cress, O'Connor, West, Allen, Wolf and Wright (2003) analyzed a dataset that included imputations generated by our method,

1. Wai Fung Chiu, Department of Statistics, Harvard University, One Oxford Street, Cambridge MA 02138. E-mail: wfcchiu@post.harvard.edu; Recai M. Yucel, Department of Biostatistics and Epidemiology, 408 Arnold House, School of Public Health and Health Sciences, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304. E-mail: yucel@schoolph.umass.edu; Elaine Zanutto, The Wharton School, University of Pennsylvania, 466 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia PA 19104. E-mail: zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA 02115. E-mail: zaslavsky@hcp.med.harvard.edu.

referring to Rubin and Zanutto (2001) and a preliminary version of this paper that appeared in a proceedings publication (Chiu, Yucel, Zanutto and Zaslavsky 2001). This paper is the first comprehensive publication of our methodology and the first published report that describes an application of Rubin and Zanutto's method to real data.

The organization for the rest of this paper is as follows. Section 2 summarizes Rubin and Zanutto's method and gives a general description of our method. Section 3 outlines the application of our method to the colorectal cancer study. Section 4 illustrates in a simulation study the performance of our method relative to three other commonly-used nonresponse adjustment methods.

2. Imputation Methodology

This section will begin with a summary of Rubin and Zanutto's method, followed by a general description of our method that includes a discussion on out-of-sample versus within-sample matching, the details of the modeling and multiply-imputing tasks, and an analysis of efficiency as a function of the number of matched cases used.

2.1 Matching, Modeling and Multiply Imputing

Rubin and Zanutto (2001) proposed a method called "matching, modeling, and multiply imputing" (MMM) that uses matched substitutes to help generate multiple imputations for nonrespondents in sample surveys, without requiring that substitutes be perfect replacements for the nonrespondents. Matched substitutes are responding survey units chosen to match the nonrespondents on one or more "matching covariates" – variables that are available prior to the survey and are convenient for matching but not necessarily for modeling. As a result of matching, nonrespondents and their substitutes may share similar values in their "field covariates" – variables that are only implicitly observed and are therefore not available for data analysis. "Modeling covariates" are variables that can be included in statistical models to adjust for observed differences between nonrespondents and their substitutes, but that may not be available or used for matching. The essence of MMM is that both matching and modeling covariates are used, in the context of proper multiple imputation (Little and Rubin 1987, pages 258 – 259 and references therein).

Consider a simple example where age and address covariates are available for all units in a population prior to sampling. Finding substitutes matching nonrespondents with respect to both age and address may be difficult. An alternative is to match only on address (*e.g.*, choosing a neighbor to be a substitute) and adjust for systematic age

differences between nonrespondents and matched substitutes through statistical modeling. If neighboring households were chosen as matched substitutes for nonresponding households, the substitutes and nonrespondents might have similar socioeconomic contexts (*e.g.*, levels of crime, access to public transportation, *etc.*) even though these characteristics might have not been recorded. In this example, address is a matching covariate, age is a modeling covariate, and the contextual socioeconomic characteristics are field covariates.

In summary, MMM (i) chooses matched substitutes for nonrespondents and some respondents based on matching covariates, (ii) uses modeling covariates to fit a model estimating the systematic differences in responses between pairs of respondents and substitutes, (iii) multiply-imputes the unobserved values using the model in (ii) under the assumption that the same relationship holds between pairs of nonrespondents and substitutes, and (iv) discards all matched substitutes after imputation.

2.2 Out-of-Sample Versus Within-Sample Matching

Matched cases may be obtained from out-of-sample data or within-sample data. In the Rubin and Zanutto approach, matched substitutes are obtained from out-of-sample data *after* the missingness is detected. Their description emphasizes that the matched substitutes must be discarded after imputation since including such additional cases in inferences would modify the sample design by adding extra cases in the "blocks" that contain unobserved data. Matched cases are considered within-sample data if they are obtained from the database that is available *before* imputing or even finding out which records in the database have unobserved variables. As far as the overall inferential goals are concerned, these matched cases are not additional cases, but are part of the original data collection, and therefore will be included in scientific analyses.

Assuming within-sample matching, we treat the ungeocodable records as nonrespondents and the geocodable records as respondents. For each ungeocodable record, a given number of matched cases are randomly chosen from a pool of geocodable records within the same small geographical area (*e.g.*, zip code, which is a postal delivery code usually representing an area served by a single main US post office). Similarly, the same number of matched cases are also chosen for each of the randomly sampled geocodable records (see Rubin and Zanutto (2001) for recommendations on the size of such a sample relative to the total number of ungeocodable records in a given dataset). If more matched cases were needed than those are available in the same small area, the selection pool would be extended to the "nearest" geographical areas until the required number of matched cases was achieved.

All matched cases in the colorectal cancer study came from the same cancer database. In general, matched cases need not be drawn from the same population in which the nonrespondents and respondents originated. For example, matched cases for colorectal cancer records can be obtained from a general population of cancer patients, and a model can then be fitted to correct for systematic differences. Note that, with matched cases from a more similar population, stronger models can be built with more covariates. In our example, since we used other patients with the same cancer type, relationships to treatment process and outcome variables are likely to be consistent.

2.3 Modeling and Multiply-imputing

A simple example of our method is given here to convey the basic idea; in practice, more complex models may often be required. Suppose the following relationship holds in the population,

$$y_{ik} = \mathbf{x}_{ik}^T \boldsymbol{\beta} + \delta_i + \varepsilon_{ik}, \quad (1)$$

where i indexes small geographical area, k indexes unit within area, and y_{ik} and \mathbf{x}_{ik} are respectively the response and the characteristics of the k^{th} unit in geographical area i . This model includes a regression prediction $\mathbf{x}_{ik}^T \boldsymbol{\beta}$, a small-area effect δ_i , and a unit-specific residual ε_{ik} . We assume that ε_{ik} follows some distribution F_ε with mean zero and variance σ^2 . Note that this development generalizes directly to multivariate y_{ik} .

We extend Rubin and Zanutto's method to allow more than one match in the same small area, because having several matches in small areas is possible (often convenient and inexpensive) in census data or in large administrative datasets. Rubin and Zanutto's assumption of a single match is appropriate to survey data collection that requires additional field work for each match.

The regression coefficients in equation (1) are estimated using any collection of observations with two or more records per small area to fit the regression model in which the δ_i are treated as fixed effects. With only two cases per area, $\boldsymbol{\beta}$ can instead be estimated from the within-area regression

$$(y_{i1} - y_{i2}) = (\mathbf{x}_{i1}^T - \mathbf{x}_{i2}^T) \boldsymbol{\beta} + (\varepsilon_{i1} - \varepsilon_{i2}), \quad (2)$$

where the small area effect drops out. The residuals from this regression have a symmetrical distribution with variance $2\sigma^2$.

Assuming for the moment that we have a draw from the posterior distribution of $\boldsymbol{\beta}$, we carry out the rest of this analysis conditional on that draw. Now suppose that we are interested in imputing for a new unit (indexed as $k=0$) in area i , and that we have obtained $K_i \geq 1$ matched cases for this unit. Denote the outcomes of these matched cases by

the vector $\mathbf{y}_i = (y_{i1}, \dots, y_{iK_i})^T$ and the corresponding characteristics by the matrix $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iK_i})^T$. With a flat prior for δ_i , the posterior distribution for $\delta_i | \mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\beta}$ has mean

$$\bar{y}_i - \bar{\mathbf{x}}_i^T \boldsymbol{\beta} \quad (3)$$

and variance σ^2 / K_i , where $\bar{y}_i = \sum_{k=1}^{K_i} y_{ik} / K_i$ and $\bar{\mathbf{x}}_i = \sum_{k=1}^{K_i} \mathbf{x}_{ik} / K_i$. Hence, the predictive distribution for $y_{i0} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{x}_{i0}, \boldsymbol{\beta}$ has mean

$$\bar{y}_i + (\mathbf{x}_{i0}^T - \bar{\mathbf{x}}_i^T) \boldsymbol{\beta} \quad (4)$$

and variance $(1 + 1/K_i) \sigma^2$ which is the sum of the predictive variance under the model conditional on all parameters and the posterior variance of δ_i . These statements assume that the mean of the residuals is a sufficient statistic for δ_i . This assumption is true for the normal distribution (or natural observations of any exponential family distribution); we assume it is at least approximately true for F_ε , so that we can base inferences on that mean. Note that use of a flat prior leads to overdispersed draws relative to what would be obtained with a proper prior from a hierarchical model, but is much simpler (especially in analyses with the multivariate outcomes).

An imputation for y_{i0} can be generated by first drawing σ^2 from its posterior distribution, second drawing $\boldsymbol{\beta}$ conditional on the draw of σ^2 , third computing the predictive mean in equation (4) from the draw of $\boldsymbol{\beta}$, and finally adding a residual of variance $(1 + 1/K_i) \sigma^2$ to the predictive mean. In simple surveys with $\boldsymbol{\beta}$ estimated by equation (2), the posterior distribution of $\boldsymbol{\beta}$ (conditional on σ^2 and the data) under a flat prior is approximately $N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$ where the i^{th} row of \mathbf{X} is $(\mathbf{x}_{i1}^T - \mathbf{x}_{i2}^T)$. In more complex designs, the posterior distribution of $\boldsymbol{\beta}$ can be approximated using the point estimate and sampling variance calculated under the associated design.

The residual can be obtained through modeling or sampling. Modeling involves estimating σ^2 using the residual variance of equation (1) and drawing the residual under univariate normality (see Rubin and Zanutto (2001) for the special case where only one matched case was obtained for each record) or some other parametric distribution. We refer to such an approach as **parametric MMM** (PMMM). An alternative is to randomly sample a regression residual from any area j whose residuals might be regarded as exchangeable with those from area i (Rubin 1987 pages 166–168). See also Lessler and Kalsbeek (1992, section 8.2.2.4), Kalton and Kasprzyk (1986), and Kalton (1983). Since the variance of such a residual is $[(K_j - 1)/K_j] \sigma^2$, we multiply the randomly-sampled residual by $\sqrt{[(K_i + 1)/K_i][K_j/(K_j - 1)]}$ to obtain the

correct predictive variance. We call this approach **nonparametric MMM (NpMMM)**.

In summary, our method consists of three basic steps:

1. Draw matched cases for the ungeocodable records and for some randomly sampled geocodable records;
2. Use the sampled geocodable records and their matched cases to fit equation (1) where the δ_i are treated as fixed effects, and save the residuals;
3. Repeat the following for m (usually 5 to 10) times:
 - (a) Draw σ^2 from its posterior distribution, then β conditional on the draw of σ^2 ;
 - (b) For each ungeocodable record, treat the sum of the vector of predictive means obtained from equation (4) and a vector of residuals drawn using either PMMM or NpMMM as a realization of the unobserved vector of contextual variables.

2.4 Efficiency

The efficiency of an imputation is related to the number of matched cases used. Let V_K be the predictive variance of an imputation model where K matched cases per record are used. For the model in section 2.3, $V_K = (1 + 1/K) \sigma^2$. Define efficiency as

$$E_K = \frac{V_\infty}{V_K} = \frac{\sigma^2}{(1 + 1/K) \sigma^2} = \frac{K}{K + 1}, \quad (5)$$

for any positive integer K . Efficiency increases as the number of matched cases per record increases; for example, $E_2 \approx 0.67$, $E_4 = 0.8$, $E_{10} \approx 0.91$, and $E_{20} \approx 0.95$.

Theoretically each record can have as many matched cases as permitted by available resources. In practice, the number of matched cases used often depends on the cost of matched cases and the cost of computation involved in model fitting. In our method, the cost of computation for each added matched case per record is negligible. In the colorectal cancer study, while the matched cases were free, the ability to do the imputation based on a limited number of matched cases was crucial because confidentiality restrictions prevented investigators from using the entire dataset in modeling with zip codes (even in a coded form) attached. For illustrative purposes, we will use two matched cases per record in subsequent analyses.

3. Application: The Colorectal Cancer Study

The colorectal cancer database has a total of 50,740 patient records, of which approximately 3.3% are ungeocodable. Among these, about half have P.O. box addresses (often in a rural area), and the rest are mistyped

addresses or addresses from newly developed areas that are not in address databases. In a study of factors predicting provision of chemotherapy for colorectal cancer patients, investigators believed that the following three census block-group means would be useful contextual variables:

- Y_1 = median household income,
- Y_2 = percent with no high school diploma, and
- Y_3 = percent below poverty level.

These variables were observed in geocodable records but unobserved in ungeocodable records. The task was to generate multiple imputations for the unobserved census variables using the methods in section 2.

Each of the block-group means was reported in the census data for six race/ethnic groups, and the scientific analyses used only the set of block-group means corresponding to the race/ethnicity of each patient. For imputations used in Ayanian *et al.* (2003), we therefore fitted six separate models to impute all $18(6 \times 3)$ values for each ungeocodable patient and then selected the three variables pertinent to each patient; joint distributions for different race/ethnic groups were not important because each imputation only used values for a single group. An alternative would have been to use race as a matching variable, but this would have forced us to seek some matches at a much greater distance geographically, diluting the predictive value of the geographical match.

For expository purposes, we assume henceforth that only the block-group mean corresponding to the race of each respondent is available, but not the means corresponding to the other five races that are available simultaneously in the census data. This is more typical of data that would be collected directly from the respondent, where the race variable itself (as a modeling variable) is quite predictive because income data for people of different races reflect differences in income associated with race.

3.1 Matching and the Dataset

The addresses of over 90% of ungeocodable records have zip codes. Zip code was therefore chosen as a matching covariate. A simple diagnostic for its usefulness appears in section 3.2. The numerical sequence of zip codes does not always correspond to neighborhood distance relationships. For example, Cambridge, Massachusetts has a 02138 post office that also uses the 02238 zip code for mailboxes, and in nearby Boston there is a 02215 zip code that was carved out of the 02115 area. Instead of using the numerical sequence of zip codes, the distances between zip codes were computed based on latitudes and longitudes of their main post offices, under the assumption that two zip codes were closest to each other if their main post offices were closest to each other.

The colorectal cancer database has 1,696 ungeocodable records. The same number ($n^* = 1,696$) of geocodable records was randomly selected from the same database. For each of these 3,392 records, two matched geocodable cases were randomly chosen from its own zip code or (if necessary) neighboring zip codes. This created a dataset with $3,392 \times 3 = 10,176$ records. Note that n^* was a convenient choice, because the data were free. In general, the choice of n^* could affect both the total cost and the precision of the estimates. Both the randomly selected geocodable records and the matched cases were within-sample data and hence were retained in the analyses for Ayanian *et al.* (2003). We asked the cancer registry for these cases only because for confidentiality purposes we could not do the matching ourselves with the data (for the same cases) that we had in hand.

The modeling covariates used in the imputation model were the eight administrative-record variables: age, sex, race, marital status, cancer stage, chemotherapy treatment, cancer type and radiotherapy treatment, and category of treating hospital's American College of Surgeons accreditation as of 1999 (ACOS99). These variables are observed for all 10,176 records included in the imputation model. (Some of these variables are predictors and some are outcomes in the scientific models of the main analyses, but the distinction is irrelevant for imputation.) The census mean values Y_1 , Y_2 and Y_3 are observed in geocodable records, but not in ungeocodable records. These variables were treated as outcome variables of the imputation model in section 2.3. The data structure is represented by Table 1.

Table 1
Structure of Data Used in Imputation for the
Colorectal Cancer Study

Data*	Eight Modeling Covariates				Census Variables		
	Age	Sex	...	ACOS99	Y_1	Y_2	Y_3
Ungeocodable	✓	✓	...	✓	?	?	?
First Match	✓	✓	...	✓	✓	✓	✓
Second Match	✓	✓	...	✓	✓	✓	✓
Geocodable	✓	✓	...	✓	✓	✓	✓
First Match	✓	✓	...	✓	✓	✓	✓
Second Match	✓	✓	...	✓	✓	✓	✓

* There were 1,696 records in each of the six types of data.
✓ = observed ? = unobserved

Before we fitted the model, the percentage outcomes y_2 and y_3 were transformed using the scaled-logit function:

$$\log \left(\frac{(y - a)/(b - a)}{1 - (y - a)/(b - a)} \right), \tag{6}$$

with $a = -0.5$ and $b = 100.5$ so that after imputations the inverse transformation with rounding to the nearest integer

would yield imputed values between 0 and 100 inclusive (Schafer 1999). Similarly, a log-transformation was applied to the income outcome y_1 so that the imputed incomes would be nonnegative. Note that the distributions of the transformed variables are closer to normality than they are on the original scale (Schafer 1997). To keep notation simple, we redefine y_1 , y_2 and y_3 as their transformed versions.

3.2 Preliminary Diagnostics

A simple diagnostic test for the usefulness of the matching covariates is to compare the adjusted R^2 for the regression models predicting the three census variables with only the modeling covariates, the models with only the matching covariates, and the models with both. In this application, zip code was the only matching covariate. There were 1,133 distinct zip codes (hence 1,132 dummy variables) in the 8,480 fully observed records (the geocodable records and all first and second matches). Table 2 shows the adjusted R^2 for models with only the eight modeling covariates, models with only zip code, and models with both modeling covariates and zip code. The adjusted R^2 for models with both modeling covariates and zip code are higher than the corresponding ones for models with only one of the two covariate types. Our imputation procedure uses information from both matching and modeling covariates and thus can be expected to work better than procedures using only the matching or the modeling covariates (as shown by the simulation study in section 4). Although the contribution of the modeling covariates to R^2 is relatively modest, their inclusion is important for removing systematic biases and properly representing relationships that might be important in the scientific models.

Table 2
Adjusted R^2 for Alternative Regression Models

	Only Modeling Covariates	Only Matching Covariate (Zip Code)	Both Modeling and Matching Covariates
Median household income (INC)	0.091	0.453	0.496
Percent with no high school diploma (EDU)	0.115	0.452	0.503
Percent below poverty level (POV)	0.047	0.327	0.343
Model degrees of freedom ^(a)	26 ^(b)	1,133	1,158
Sample sizes	8,480	8,480	8,480
Residual degrees of freedom	8,454	7,347	7,322

- (a) With intercept.
(b) The modeling covariates are age, sex (2 levels), race (6 levels), marital status (6 levels), cancer stage (6 levels), chemotherapy treatment (2 levels), cancer type and radiotherapy treatment (3 levels), and category of treating hospital's American College of Surgeons accreditation as of 1999 (6 levels).

To determine whether a multivariate model was needed, we fitted a multivariate-outcome regression model with both

modeling covariates and zip code. The estimated correlations between the residuals were: $r_{12} \approx -0.194$, $r_{13} \approx -0.297$, and $r_{23} \approx 0.357$, where “variable 1” is median household income, “variable 2” is percent with no high school diploma, and “variable 3” is percent below poverty level. These estimates were significantly different from zero, which therefore indicated that multivariate versions of the methods in section 2.3 should be used to generate imputations.

3.3 Multiple Imputation Results and Comparisons

Imputations under NpMMM were used in the study of factors predicting provision of chemotherapy for colorectal cancer patients (Ayanian *et al.* 2003). Their model included three indicator variables for ranges of contextual income, together with 21 other variables representing patient and hospital characteristics. The multiple imputation analysis shows that the information loss due to missing information is always less than 0.1%, which is much smaller than the fraction of ungeocodable records (3.3%). As expected, the largest fractions of missing information appeared for the income variables. The scientific results in Ayanian *et al.* (2003) would not have changed dramatically if the incomplete cases had been dropped. In this type of research, however, every case is precious and expensive, and saving the 3.3% with missing data was a contribution to the study.

For comparison, variances of parameters under the complete-case analysis were on the average 4.0% larger than those under multiple imputation analysis. Such percentage differences are close to the fraction of incomplete cases deleted for this analysis. When the imputations generated by our method were included in the scientific analysis, the precision of the estimate of the “rural” effect was dramatically improved (using only the complete cases led to 41.6% increase in variance), due to the concentration of ungeocodable records in rural areas (21.6% of rural records are ungeocodable, but only 3.1% of nonrural records are ungeocodable).

4. A Simulation Study

This simulation study compares performance of our new method with three other commonly-used nonresponse adjustment methods. The population of this study was the 1,696 fully observed triples – the 1,696 geocodable records and the corresponding first and second matches (one row from each of the last three horizontal blocks in Table 1) – or 5,088 observations. For simplicity, we assumed that the triples were from distinct zip codes (clusters), hence $i = 1, 2, \dots, I = 1,696$. Each cluster i contained three units ($u = 1, 2, 3$), and the record of each unit consisted of \mathbf{x}_{iu} (the covariates) and \mathbf{y}_{iu} (the census variables).

4.1 Simulated Data and Response Mechanism

Assuming that the design was cluster sampling with sample size 800, we drew random samples of 800 clusters. For each random sample, about half of the 800 clusters were randomly selected to have an ungeocodable record in which the census variables were unobserved, with the probability of missingness depending on an individual’s race and on the mean income of the cluster (zip code). We simulated missingness under a multinomial logit model where the outcomes are: nothing unobserved ($w_{i0} = 1$), y_{i1} unobserved ($w_{i1} = 1$), y_{i2} unobserved ($w_{i2} = 1$), and y_{i3} unobserved ($w_{i3} = 1$). Specifically, for each $i = 1, 2, \dots, I$, let $z_{iu} = 0$ and

$$z_{iu} = a + b \times I(\text{unit } iu \text{ is White}) \\ + c \times (\text{mean income in zip code } i) \quad (7)$$

where $u = 1, 2, 3$. Then

$$\Pr(w_{iu} = 1) = \exp(z_{iu}) / \sum_{u=0}^3 \exp(z_{iu}) \\ \text{for } u = 0, 1, 2, 3. \quad (8)$$

The results of this simulation study were based on datasets generated by the mechanism with $a = -1$, $b = 11$ and $c = 0.0003$, which made about 17% of the units in a random sample ungeocodable, with probability of geocoding positively related to White race and higher block-level income. The task was to use the random sample to estimate $\bar{\mathbf{y}}$, the mean values of the population (1,696 clusters).

The simulation conditions described in the preceding paragraphs were designed to give a stringent test of the procedure and alternatives by exaggerating the impact of unobserved data and making the missingness strongly related to characteristics both of the individual and of the area. We were not attempting to simulate the exact conditions of the application in section 3 but rather to use an artificial population with similar distributions to those in the real population to illustrate the workings of our method and its competitors.

4.2 Inferential Methods and Measures of Performance

Preliminary results indicated that the performance of PMMM and NpMMM is similar; NpMMM is, however, simpler (especially in analyses with multivariate outcomes), because the method does not require explicit parametric modeling of the residual variance. Our simulations compared performance of NpMMM (using two matched cases per record) with three other commonly-used nonresponse adjustment methods:

1. Complete-case Method (CCM)

The population means are estimated from all geocodable units of a random sample.

2. Substitute Single Imputation (SSI)

This is the traditional use of substitutes. The unobserved census variables of each ungeocodable unit are replaced by the values of the census variables of a randomly selected unit from the same cluster. The resulting sample is treated as if there had been no ungeocodable unit; all 800 clusters in such a sample are used for estimating the population means.

3. Multivariate Normal Multiple Imputation (MNMI)

This method uses only one randomly selected unit from each of the fully observed clusters in a random sample to fit the multivariate normal linear regression

$$\mathbf{y}_i^T \sim N(\boldsymbol{\beta}_0^T + \mathbf{x}_i^T \mathbf{B}, \boldsymbol{\Sigma}),$$

with a noninformative prior on the parameters. The model is then used to create m sets of multiple imputations for the unobserved census variables using a direct multivariate generalization of the algorithm given by Rubin (1987, page 167).

Note that CCM uses *neither* matching nor modeling covariates, SSI uses *only the matching covariate* (zip code), MNMI uses *only the modeling covariates*, and NpMMM uses *both* the matching covariate and the modeling covariates.

The CCM and SSI data are analyzed by the usual complete-data method which estimates the population mean from the data with the appropriate estimator for cluster sampling from a finite population, including the finite population correction (Cochran 1977, Chapters 9–10). Both MNMI and NpMMM produce m sets of complete data, each of which is analyzed by the same complete-data method used for the CCM and SSI data; the m sets of point and variance estimates are then combined using the multiple imputation combination rule (Rubin 1987; Schafer 1997, pages 108–110).

For each simulation $t \in \{1, 2, \dots, T\}$, we denote the point estimates from the four methods by $\bar{\mathbf{y}}_{CC}(t)$, $\bar{\mathbf{y}}_{SS}(t)$, $\bar{\mathbf{y}}_{MN}(t)$, and $\bar{\mathbf{y}}_{Np}(t)$, and the means of these quantities across simulations are written as $\bar{\mathbf{y}}_{CC}$, $\bar{\mathbf{y}}_{SS}$, $\bar{\mathbf{y}}_{MN}$, and $\bar{\mathbf{y}}_{Np}$. Performance evaluation of the four nonresponse adjustment methods will be based on three measures:

- Percent reduction in the average bias of an estimator relative to the average bias of the CCM estimator.** Denote the average bias of an estimator by \bar{b}_E . Then

$$\bar{b}_E = \bar{y}_E - \bar{y},$$

where $E \in \{CC, SS, MN, Np\}$. We define the percent reduction in the average bias of an estimator relative to the average bias of the CCM estimator as

$$R(\bar{b}_E, \bar{b}_{CC}) = \frac{|\bar{b}_{CC}| - |\bar{b}_E|}{|\bar{b}_{CC}|},$$

where \bar{b}_E is an element of $\bar{\mathbf{b}}_E$ and \bar{b}_{CC} is the corresponding element in $\bar{\mathbf{b}}_{CC}$. By definition, $R(\bar{b}_{CC}, \bar{b}_{CC})$ is zero.

- Estimated coverage of the nominal 95% confidence intervals for $\bar{\mathbf{y}}$.** Intervals produced by the CCM or SSI estimates were constructed under appropriate t -distributions. For intervals associated with the MNMI or NpMMM estimates, we followed the procedure outlined in Schafer (1997, pages 109–110) and replaced the degrees of freedom ν with the updated version of Barnard and Rubin (1999).
- Estimated fraction of missing information about $\bar{\mathbf{y}}$.** For each of MNMI and NpMMM, we computed $\hat{\lambda}$, an estimate of the fraction of missing information about $\bar{\mathbf{y}}$ (see Barnard and Rubin (1999) for the most recent expression).

4.3 Results

The simulation procedure was implemented 2,000 times, and $m=10$ was used for MNMI and NpMMM. The mean values of the census variables in the population were $\bar{\mathbf{y}} = (40,642, 21.65, 9.55)^T$. The average bias of the CCM estimator was $\bar{\mathbf{b}}_{CCM} = (-5,405, -3.97, -1.79)^T$. Other results are summarized in Table 3. NpMMM achieved large percent reductions in relative average bias (95.0% to 99.5%). SSI reduced biases more than MNMI, because the matching covariate (zip code) was much more informative than the set of modeling covariates (section 3.2). Since the response mechanism was *nonignorable* (the response probabilities depended partly on income), the poor performance of MNMI, which did not use the geographical information to help predict income, was expected. Note that MNMI is biased, and the bias is large enough so that with the sample size considered in this paper the confidence intervals never covered the hypothetical population values.

Under MNMI and NpMMM, the percent of missing information was much less than the average percent of unobserved data. The percent of missing information was smaller under NpMMM than under MNMI. Only NpMMM produced well calibrated intervals with correct coverage. In summary, NpMMM combines the best features of the other two methods – close-to-nominal coverage and less missing information.

Table 3

Simulation Results^(a): Bias Reduction, Coverage, and Fraction of Missing Information

Measure	Mean	Method		
		NpMMM	MNMI	SSI
Percent bias	INC	99.5	44.6	95.2
Reduction	EDU	95.0	40.6	83.7
$100R(\bar{b}_E, \bar{b}_{CCM})^{(b)}$	POV	96.8	32.6	80.3
Estimated	INC	95.1	0.00	89.8
Coverage of the	EDU	94.8	0.00	65.7
95% CIs ^(c)	POV	95.2	0.00	66.0
100× Estimated	INC	1.00	9.92	
fraction of missing	EDU	0.05	0.07	
information $\hat{\lambda}^{(d)}$	POV	0.07	0.08	

- (a) Based on 2,000 replications and $m = 10$.
(b) By definition, $100R(\bar{b}_{CCM}, \bar{b}_{CCM}) = 0$.
(c) Results for the CCM estimates were all zeros.
(d) The average percent of unobserved data was approximately 17%.

5. Conclusion

This work extends Rubin and Zanutto (2001) in two respects. First, our method allows more than one matched case per record. We show theoretically that the efficiency of an imputation increases as the number of matched cases per record increases. When the cost of matched cases is relatively low, our method offers an option where information of more than one matched case per record is used to help fit imputation models at a negligible computational expense. Second, NpMMM does not require explicit parametric modeling of residual variance(s), hence simplifying the modeling task (especially for analyses with multivariate outcomes). This nonparametric approach makes it feasible to apply our method to datasets with complex model structures. In a simulation study, NpMMM estimates achieved substantial bias reductions, and NpMMM produced confidence intervals with correct coverage.

Although we have focused on geographically-based matching to complete unobserved geographically-linked variables, the procedures described in this paper can be generalized to other matching variables. For example, to impute clinical variables, it might be more appropriate to match to another patient in the same hospital, if clinical characteristics and therapies are likely to be more strongly associated with the hospital than with the geographic location of the patient’s residence.

Acknowledgements

This research was supported in part by the Bureau of the Census through a contract with the National Opinion Research Center and Datametrics, Inc., and by a grant from the Agency for Healthcare Research and Quality (AHRQ) and the National Cancer Institute (HS09869). The authors thank John Z. Ayanian for leadership of the Quality of Cancer Care research project, Mark Allen and Robert Wolf for preparation of data, Bill Wright for his support to this research, and the associated editor and two anonymous referees for their helpful comments.

References

Ayanian, J.Z., Zaslavsky, A.M., Fuchs, C.S., Guadagnoli, E., Creech, C.M., Cress, R.D., O’connor, L.C., West, D.W., Allen, M.E., Wolf, R.E. and Wright, W.E. (2003). Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. *Journal of Clinical Oncology*, 21, 1293-1300.

Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955.

Chiu, W.F., Yucel, R.M., Zanutto, E. and Zaslavsky, A.M. (2001). Using matched substitutes to improve imputations for geographically linked databases. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Kalton, G. (1983). *Compensating for Missing Survey Data*. Research Report Series, Ann Arbor, MI: Institute for Social Research.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Krieger, N., Williams, D. and Andmoss, N. (1997). Measuring social class in U.S. public health research: Concepts, methodologies, and guidelines. *Annual Review of Public Health*, 18, 341-378.

Lessler, J.T., and Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley & Sons, Inc.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B., and Zanutto, E. (2001). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. In *Survey Nonresponse*, (Eds. R. Groves, R. Little and J. Eltinge), New York: John Wiley & Sons, Inc., 389-402.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT available at <http://www.stat.psu.edu/~jls/misoftwa.html>.

Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES Data

Balagobin Nandram and Jai Won Choi¹

Abstract

We use hierarchical Bayesian models to analyze body mass index (BMI) data of children and adolescents with nonignorable nonresponse from the Third National Health and Nutrition Examination Survey (NHANES III). Our objective is to predict the finite population mean BMI and the proportion of respondents for domains formed by age, race and sex (covariates in the regression models) in each of thirty five large counties, accounting for the nonrespondents. Markov chain Monte Carlo methods are used to fit the models (two selection and two pattern mixture) to the NHANES III BMI data. Using a deviance measure and a cross-validation study, we show that the nonignorable selection model is the best among the four models. We also show that inference about BMI is not too sensitive to the model choice. An improvement is obtained by including a spline regression into the selection model to reflect changes in the relationship between BMI and age.

Key Words: Cross-validation; Deviance; Metropolis-Hastings sampler; Normal-logistic regression model; Spline regression model.

1. Introduction

The National Health and Nutrition Examination Survey (NHANES III) is one of the surveys used by the National Center for Health Statistics (NCHS) to assess the health of the U.S. population. One of the variables in this survey is body mass index (BMI), and the World Health Organization has used BMI to define overweight and obesity. Under ignorability estimators from the NHANES III data are biased because there are many nonrespondents, and the main issue we address here is that nonresponse should not be ignored because respondents and nonrespondents may differ. The purpose of this work is to predict the finite population mean BMI for children and adolescents, post-stratified by county for each domain formed by age, race and sex and to investigate what adjustment needs to be made for nonignorable nonresponse. Our approach is to fit several hierarchical Bayesian models to accommodate the nonresponse mechanism.

Recently, several articles have been written about overweight and obesity. In outlining the first national plan of action for overweight and obesity, the Surgeon General called for sweeping changes in schools, restaurants, workplaces and communities to help combat the growing epidemic of Americans who are overweight or obese. He said that the obesity report "Is not about esthetics and it's not about appearances. We're talking about health." As noted by Squires (2001) "Health care costs for overweight and obesity total an estimated \$117 billion annually." Overweight children often become overweight in adulthood,

and overweight in adulthood is a health risk (Wright, Parker, Lamont and Craft 2001). In a very interesting article, using NHANES data Ogden, Flegal, Carroll and Johnson (2002) describe the most recent national estimates of the prevalence and trends in overweight among U.S. children and adolescents. Based on a limited analysis they conclude "The prevalence of overweight among children in the United States is continuing to increase especially among Mexican-American and non-Hispanic black adolescents." Several disorders have been linked to overweight in childhood. A potential increase in type 2 diabetes mellitus is related to the increase in overweight among children (Fagot-Campagna 2000); so are cardiovascular risk factor, high cholesterol levels, and abnormal glucose levels (Dietz 1998). Thus, it would be helpful to study the BMIs for children and adolescents using methods that can provide accurate adjustment for nonresponse and better measure of precision.

Letting x denote covariates and y the response variable, Rubin (1987) and Little and Rubin (1987) describe three types of missing-data mechanism. These types differ according to whether the probability of response (a) is independent of x and y (b) depends on x but not on y and (c) depends on the y and possibly x . The missing data are missing completely at random (MCAR) in (a), missing at random (MAR) in (b) and one may say that the data are missing not at random (MNAR) in (c). Models for MCAR and MAR missing-data mechanisms are called ignorable if the parameters of the dependent variable and the response are distinct (Rubin 1976). Models for MNAR missing-data mechanisms are called nonignorable.

1. Balagobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280. E-mail: balnan@wpi.edu; Jai Won Choi, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. E-mail: jwc7@cdc.gov.

Nonresponse models can be classified very broadly into selection and pattern mixture models (e.g., see Little and Rubin 1987). Let $[y]$ and $[r]$ denote respectively the density function of the response variable y , and the response indicator r , with obvious notations for the joint and conditional densities. Then the selection model specifies that $[y, r] = [r|y][y]$ and the pattern mixture model specifies $[y, r] = [y|r][r]$. The selection approach was developed to study sample selection problems (e.g., Heckman 1976 and Olson 1980). While the two models have the same joint density, in practice the components $[r|y]$ and $[y]$ for the selection model, and $[y|r]$ and $[r]$ for the pattern mixture model are specified. Thus, these models may differ.

Thus, we use two nonignorable nonresponse models, a selection model and a pattern mixture model, to analyze the NHANES III data. Each model is used in the hierarchical Bayesian frame work for our nonignorable nonresponse problem, and to study sensitivity to model choice the results are compared. In the selection model, the response propensity is related to BMI only, and then the model on BMI has a linear model on age, race, sex and the interaction of race and sex. In the pattern mixture model, the propensity to respond is related to age, race and sex (not BMI), and the model on BMI has two closely related linear forms on age, race, sex and the interaction of race and sex. These two models hold for the entire population. The BMI values of the nonrespondents and the nonsampled individuals are predicted from each model. We prefer the selection model because we can incorporate the structure in the NHANES III data, and based on statistical arguments this turns out to be true.

Greenlees, Reece and Zieschang (1982) developed a normal-logistic regression model for imputing missing values when the probability of response depends upon the variable being imputed. They applied the model to data on wages and salary in the Current Population Survey (CPS) data on wages. David, Little, Samuel and Triest (1986) compared the CPS hot deck method and the normal-logistic regression model to wages and salary from a similar data set, and they found very little difference between the two methods. We note that the normal-logistic regression model is a nonignorable nonresponse selection model, but it does not account for clustering. To accommodate clustering within counties in the NHANES III data, it is natural to start with the normal-logistic model.

Our hierarchical Bayesian selection model has a special structure. In NHANES III the propensity to respond increases with age (race and sex play a minor role), and doctors believe that obese individuals tend not to turn up for the physical examination. Thus, given the BMI values, like Greenlees *et al.* (1982) the response indicators follow a

logistic regression model with the logarithm of the BMI values being the covariate. In turn, the logarithms of the BMI values are distributed according to a linear model in which the covariates are age, race and sex. This is the most important information we incorporate into the selection model. In addition, unlike Greenlees *et al.* (1982) our model includes clustering effects to account for heterogeneity among counties through the response indicators and the BMI values. Here each county has its own set of parameters, and there is a common distribution over these sets of parameters. This is also an important prior information we incorporate into our model, and it is one of the attractive features of the hierarchical Bayesian methodology.

In the Bayesian approach, the main difficulty is formulating the relationship between the respondents and non-respondents. This latter issue can be accommodated within the selection approach through the normal-logistic structure. We also consider a hierarchical Bayes model within the pattern mixture approach. The pattern mixture model is a useful alternative to study sensitivity to the assumption in the selection model. To assess the assumption of non-ignorable nonresponse, we also consider special cases of the selection and pattern mixture models to obtain two ignorable models. We found that a fifth model is required, in which we extend our selection model to a spline regression model to accommodate the dynamic relation between BMI and age.

Nandram, Han and Choi (2002) developed a methodology to analyze the BMI data by age, race and sex when BMI is categorized into three intervals. This is a multinomial extension of the nonresponse nonignorable analysis of Stasny (1991) for binary data. This methodology applies generally to any number of cells in several areas (counties in our application). Nandram and Choi (2002 a,b) consider further extensions of the work of Stasny for binary data (*i.e.*, data from the National Health Interview Survey and the National crime survey). Here we do not categorize the BMI values, but rather we treat them in their own right as continuous values. The quantities of interest are the finite population mean BMI and the proportion of responding individuals in each domain formed by age, race, sex and county.

The rest of the paper is organized as follows. In section 2, we briefly describe the NHANES III data. In section 3, we discuss the hierarchical Bayesian models for ignorable and nonignorable nonresponse. We also describe the model fitting, model selection and assessment which use predictive deviance and cross-validation. In section 4 we describe the analysis of NHANES III BMI data. Section 5 has a description of a spline regression model and comparisons. Finally, section 6 has concluding remarks about our approach.

2. NHANES III Data

The sample design is a stratified multistage probability design which is representative of the total civilian non-institutionalized population, 2 months of age or older, in the United States. The number of sampled individuals in each age-race-sex group is known for each county. The sample size by county, age, race and sex are relatively sparse. Further details of the NHANES III sample design are available (National Center for Health Statistics 1992, 1994).

The NHANES III data collection consists of two parts: the first part is the sample selection and the interview of the members of a sampled household for their personal information, and the second part is the examination of those interviewed at the mobile examination center (MEC). The health examination has information on physical examination, tests and measurements performed by technicians, and specimen collection.

The sample was selected from households in 81 counties across the continental United States during the period from October 1988 through September 1994, but for confidentiality reasons the final data of this study came from only the 35 largest counties (from 14 states) with population at least 500,000 for selected age categories by sex and race. In this paper, we analyze public use data from these 35 counties; the demographic variables are age, race and sex, and the health indicator of our interest is body mass index (BMI), weight in kilograms divided by the square of height in meters (Kuczmarski, Carrol, Flegal and Troiano 1997). The World Health Organization (WHO Consultation of Obesity 2000) has designated an adult with BMI at least 30 as obese; overweight refers to adults with BMI in the range [25, 30). For children 1–6 years old and adolescents 7–19 years old overweight and obesity are age-dependent.

Nonresponse occurs in the interview and examination parts of the survey. The interview nonresponse arises from sampled persons who did not respond for the interview. Some of those who were already interviewed and included in the subsample for a health examination missed the examination at home or at the MEC, thereby missing all or part of the examinations. Here we do not consider the small number of individuals whose BMI values and covariates (age, race and sex) are missing (*i.e.*, unit nonresponse). For simplicity and for all practical purposes it is reasonable to include all individuals with their covariates (*i.e.*, complete data and item nonresponse) reported in our data analysis. Cohen and Duffy (2002) point out that “Health surveys are a good example, where it seems plausible that propensity to respond may be related to health.” We note also that for children and adolescents the observed nonresponse rate is about 24%. A partial reason for the nonresponse for young children is that the parents or older mothers were extremely

protective and would not allow their children to leave home for a physical examination.

We study the BMI data for four age classes (02 – 04, 05 – 09, 10 – 14, 15 – 19 years). Recalling that there are 560 ($35 \times 4 \times 2 \times 2$) domains, the sample sizes on the average are very small per domain (*e.g.*, $2,647/560 \approx 4$). Thus, there is a need to “borrow strength” from the domains. Also, the sample size is small relative to the finite population size (*e.g.*, $100 \times (2,647/6,653,738) = 0.04\%$). The prediction problem needs much computation. The observed data indicate that there is an increasing trend of BMI with age with slightly increasing variability.

NHANES III data are adjusted by multiple stages of ratio weightings to be consistent with the population; see Mohadjar, Bell and Waksberg (1994). In this ratio-method, item nonresponse adjustment is done by ratio estimation within the same adjustment class and the distributions of the respondents and nonrespondents are assumed to be same. There is a need to consider methods for handling non-ignorable nonresponse other than the ratio-adjustment method. Here we present a Bayesian method as a possible alternative for studying NHANES III nonresponse.

Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin (1996) attempted a comprehensive multiple imputation project on the NHANES III data for many variables. The purpose was to impute the nonresponse data in order to provide several data sets for public use. As one of the limitations of the project they stated “the procedure used to create missingness corresponds to a purely ignorable mechanism; the simulation provides no information on the impact of possible deviations from ignorable nonresponse.” Another limitation is that the procedure did not include geographical clustering. Our purpose is different; we do not provide imputed public-use data. Unlike Schafer *et al.* (1996), we include clustering at the county level, although there may be a need to include clustering at the household level. For the complete data there are 6,440 households. Of these households 52.1% contributed one person to the sample, 22.5% two persons, and 21.4% at least three persons. We have calculated the correlation coefficient for the BMI values based on pairing the members within households (see Rao 1973, page 199). It is 0.19 which indicates that as a first approximation the clustering within households can be ignored.

For our current application, inference is required for each age, race and sex domain within county. One standard small area estimation method is to identify each small area by a parameter, and then assume a common stochastic process over the 560 parameters. But because of the sparseness of the data, this is not desirable. Thus, our models are constructed at county level, and at the same time age, race and sex are represented as covariates. Inference is made for

each domain formed by crossing age, race and sex within county through our regression models. This is a key point in our analysis.

3. Hierarchical Bayesian Methodology

In this section we describe two Bayesian models for non-ignorable nonresponse, and we deduce two additional ignorable models as special cases. We describe the model selection and assessment for the selected model (*i.e.*, the selection model).

There are data from $\ell = 35$ counties and each county has N_i (known) individuals. We assume a probability sample of n_i individuals is taken from the i^{th} county. Let s denote the set of sampled units and ns the set of nonsampled units. Let r_{ij} for $i = 1, 2, \dots, \ell$ and $j = 1, 2, \dots, N_i$ be the response indicator ($r_{ij} = 1$ for respondents and $r_{ij} = 0$ for non-respondents) for the j^{th} individual within the i^{th} county in the population. Also, let x_{ij} be the logarithm of the BMI value. We found that the logarithm transformation gives a better representation, and we use it throughout. Note that r_{ij} and x_{ij} are all observed in the sample s but they are unknown in ns . Let $r_i = \sum_{j=1}^{N_i} r_{ij}$ (*i.e.*, r_i is the number of sampled individuals that responded in the i^{th} county).

For convenience, we express the BMI x_{ij} as $x_{i1}, x_{i2}, \dots, x_{ir_i}, x_{ir_i+1}, \dots, x_{in_i}$ in s and $x_{in_i+1}, \dots, x_{iN_i}$ in ns for county i . A key point that we note for what follows is that the r_i individuals are not necessarily random respondents from the n_i individuals randomly sampled. This is the nonresponse bias we need to address. It is clear that we need to predict the BMI value x_{ij} for (a) the nonrespondents in s and (b) the individuals in ns . Thus, for the finite population of N_i individuals, we need a Bayesian predictive inference for

$$\bar{X}_i = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i} \quad \text{and} \quad P_i = \frac{\sum_{j=1}^{N_i} r_{ij}}{N_i},$$

for $i = 1, \dots, \ell$.

Letting $\bar{x}_i^{(s,r)} = \sum_{j=1}^{r_i} x_{ij} / r_i$, $\bar{x}_i^{(s,ns)} = \sum_{j=r_i+1}^{n_i} x_{ij} / (n_i - r_i)$ and $\bar{x}_i^{(ns)} = \sum_{j=n_i+1}^{N_i} x_{ij} / (N_i - n_i)$, we note that

$$\bar{X}_i = f_i \left[g_i^{(s)} \bar{x}_i^{(s,r)} + (1 - g_i^{(s)}) \bar{x}_i^{(s,ns)} \right] + (1 - f_i) \bar{x}_i^{(ns)} \quad (1)$$

where $f_i = n_i / N_i$ and $g_i^{(s)} = r_i / n_i$. Note that while the f_i are fixed by design, the g_i and $\bar{x}_i^{(s,r)}$ are observed. Also, letting $\hat{p}_i^{(s)} = r_i / N_i$ and $\hat{p}_i^{(ns)} = (\sum_{j=n_i+1}^{N_i} r_{ij}) / (N_i - n_i)$,

$$P_i = f_i \hat{p}_i^{(s)} + (1 - f_i) \hat{p}_i^{(ns)}, \quad (2)$$

$i = 1, \dots, \ell$. We develop our hierarchical Bayesian models to perform predictive inference for quantities like (1) and (2) depending on the domain.

3.1 Competing Models

Our models have two parts, one part for the response mechanism and the other part for the distribution of BMI. These two parts are connected to form a single model under nonignorable nonresponse or ignorable nonresponse.

First, we describe the selection model. For Part 1 of this model the response depends on the BMI as follows

$$r_{ij} \mid x_{ij}, \beta_i \sim \text{Bernoulli} \left\{ \frac{e^{\beta_{0i} + \beta_{1i} x_{ij}}}{1 + e^{\beta_{0i} + \beta_{1i} x_{ij}}} \right\}, \quad (3)$$

$$\begin{aligned} &(\beta_{0i}, \beta_{1i}) \mid \theta_0, \theta_1, \sigma_1^2, \sigma_2^2, \rho_1 \\ &\stackrel{\text{iid}}{\sim} \text{BVNormal}(\theta_0, \theta_1; \sigma_1^2, \sigma_2^2, \rho_1), \end{aligned} \quad (4)$$

$$\begin{aligned} \theta &\sim N(\theta^{(0)}, \Delta^{(0)}), \sigma_1^{-2}, \sigma_2^{-2} \sim \text{Gamma}(a/2, a/2) \\ &\text{and } \rho_1 \sim \text{Uniform}(-1, 1), \end{aligned} \quad (5)$$

where $a, \theta^{(0)}$ and $\Delta^{(0)}$ are to be specified. Note that the prior densities in (5) are all jointly independent. The assumption (3) is important because it relates the response propensity to the BMI values; doctors believe that overweight and obese individuals tend not to come to the MECs for the examinations. Clustering among the counties is accommodated by (4), and it is this assumption that permits a “borrowing of strength” among the counties.

The second part of the model is about the BMI. The single most important predictor of BMI is age, with race and sex playing a relatively minor role. One possibility is to take the BMI values to be

$$x_{ij} = \mu_{ij} + \epsilon_{ij}, \quad \mu_{ij} = \alpha_{0ij} + \alpha_{1ij} a_{ij}$$

where a_{ij} denotes age and $\epsilon_{ij} \stackrel{\text{iid}}{\sim} \text{Normal}(0, \sigma_3^2)$ for $i = 1, \dots, \ell$ and $j = 1, \dots, N_i$. Also, there is a need to understand the relationship between BMI and age, race and sex. We let $z_{ij0} = 1$ for an intercept, $z_{ij1} = 1$ for non-black and $z_{ij2} = 0$ for black, $z_{ij2} = 1$ for male and $z_{ij2} = 0$ for female, $z_{ij3} = z_{ij1} z_{ij2}$ for the interaction between race and sex, and we let $\mathbf{z}'_{ij} = (z_{ij0}, z_{ij1}, z_{ij2}, z_{ij3})$. Then, for a regression of BMI on age adjusting for race and sex, letting $\mathbf{a}'_1 = (\alpha_{01}, \alpha_{02}, \alpha_{03}, \alpha_{04})$ and $\mathbf{a}'_2 = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14})$, we take $\alpha_{0ij} = \mathbf{z}'_{ij} \mathbf{a}_1 + v_{0i}$ and $\alpha_{1ij} = \mathbf{z}'_{ij} \mathbf{a}_2 + v_{1i}$ to get

$$\mu_{ij} = (\mathbf{z}'_{ij} \mathbf{a}_1 + v_{0i}) + (\mathbf{z}'_{ij} \mathbf{a}_2 + v_{1i}) a_{ij}$$

where v_{0i} and v_{1i} are random effects centered at zero with bivariate normal distribution shown below for each model.

Thus, in Part 2 of the selection model, we assume

$$\begin{aligned} x_{ij} &= (\mathbf{z}'_{ij} \mathbf{a}_1 + v_{0i}) + (\mathbf{z}'_{ij} \mathbf{a}_2 + v_{1i}) a_{ij} + e_{ij} \\ &\text{and } e_{ij} \mid \sigma_3^2 \stackrel{\text{iid}}{\sim} \text{Normal}(0, \sigma_3^2), \end{aligned} \quad (6)$$

$$(v_{0i}, v_{1i}) | \sigma_4^2, \sigma_5^2, \rho_2 \stackrel{\text{iid}}{\sim} \text{BVNormal}(0, 0; \sigma_4^2, \sigma_5^2, \rho_2). \quad (7)$$

Again, clustering among the counties is accommodated by (7), and it is this assumption that permits a “borrowing of strength” among the counties. For this part of the model, we use the prior

$$\begin{aligned} \alpha_1 &\sim \text{Normal}(\alpha_2^{(0)}, \Delta_2^{(0)}) \text{ and } \alpha_2 \sim \text{Normal}(\alpha_3^{(0)}, \Delta_3^{(0)}), \\ \sigma_3^{-2}, \sigma_4^{-2}, \sigma_5^{-2} &\stackrel{\text{iid}}{\sim} \text{Gamma}(a/2, a/2) \text{ and} \\ \rho_2 &\stackrel{\text{iid}}{\sim} \text{Uniform}(-1, 1) \end{aligned} \quad (8)$$

where $a, \alpha_k^{(0)}$ and $\Delta_k^{(0)}, k=1, 2$ are to be specified. Note that the prior densities in (8) are all jointly independent.

The nonignorable nonresponse pattern mixture model is presented in Appendix A. We have included race, sex and their interaction in the response part of the model, although these turn out to be unnecessary. The difference between the respondents and the nonrespondents in the pattern mixture model is that the intercepts in the regression vary with counties for the respondents but not for the nonrespondents; other parameters are the same. In this way we are able to “center” the nonignorable nonresponse model on the ignorable nonresponse model with some variation; see Nandram and Choi (2002 a) for a similar idea. We need to do so because the parameters become unidentifiable if substantial difference between the respondents and the nonrespondents is assumed in the nonignorable nonresponse model without the scientific knowledge. While we have used random effects to discriminate between the respondents and the nonrespondents, the parameters providing systematic difference between the respondents and nonrespondents in model of Rubin (1977), are not identifiable. Note that while in the pattern mixture model in (A.4) there are two specifications/patterns for x_{ij} (i.e., $r_{ij}=0$ and $r_{ij}=1$), but in the selection model there is a single specification.

We show how to specify parameters like $\theta^{(0)}, \Delta^{(0)}, \alpha_k^{(0)}, \Delta_k^{(0)}, k=1, 2$ in Appendix C. For a proper diffuse prior we choose a to be a value like 0.002. One can also use a shrinkage prior on σ_1^{-2} and σ_2^{-2} (see Natarajan and Kass 2000; and Daniels 1999). But this is not necessary in the hierarchical model.

It is an attractive property of the hierarchical Bayesian model that it introduces correlation among the variables. For example, in the selection model, (4) and (7) introduce a correlation among the r_{ij} and the x_{ij} , respectively. This is the clustering effect within the areas. Such an effect can be obtained directly, but it will not be as simple as in a hierarchical model. A further benefit of the hierarchical model is that it takes care of extraneous variations among

the areas; this is intimately connected to the cluster effect. Yet another benefit is that there is robustness in the model specifications at deeper levels beyond the sampling process (e.g., inference with (5) and (8) is fairly robust to moderate perturbations of the specifications of the hyperparameters). We have found this empirically here and elsewhere.

We obtain an ignorable nonresponse selection model by setting $\beta_{1i}=0$ for all counties with appropriate adjustment in the selection model. For an ignorable nonresponse pattern mixture model we set $x_{ij} = (z'_{ij} \alpha_1 + v_{0i}) + (z'_{ij} \alpha_2 + v_{1i}) a_{ij} + \epsilon_{ij}$ for both values of the r_{ij} .

3.2 Model Fitting

In this section we describe how to use the Metropolis-Hastings sampler to fit the models. We also use a deviance measure to select the best model among our four models. Then, we use a cross-validation analysis to assess the goodness of fit of the selected model, and because the same general principle applies to the four models, we describe model fitting for the selection model only.

Thus, we now combine the model for the response mechanism and the model for the BMI values to obtain the joint posterior density of all the parameters. The x_{ij} for $j=r_i+1, \dots, n_i, i=1, \dots, \ell$ are unknown; that is, they are latent variables. We denote these latent variables by $\mathbf{x}^{(s, nr)}$ and the observed data are denoted by \mathbf{x}^{obs} . Using Bayes' theorem to combine the likelihood function and joint prior distribution, we obtain the joint posterior density which, apart from the normalization constant, is $p(\mathbf{x}^{(s, nr)}, \sigma^2, \alpha, \beta, \nu, \theta, \rho_1, \rho_2 | \mathbf{x}^{(s, r)})$ and is given in (B.1) in Appendix B.

The posterior density in (B.1) is complex, so we used Markov chain Monte Carlo (MCMC) methods to draw samples from it. Specifically, we used the Metropolis-Hastings sampler (see Chib and Greenberg 1995 for a pedagogical discussion). We also used the trace plots and autocorrelation diagnostics reviewed by Cowles and Carlin (1996) to study convergence and we used the suggestion of Gelman, Roberts and Gilks (1996) to monitor the jumping probability in each Metropolis step in our algorithm. In performing the computation, centering the BMI values help in achieving convergence (see Gelfand, Sahu and Carlin 1995). However, this is not quite a straightforward task because centering in the logistic regression affects the BMI part of the model as well.

We obtained a sample of 1,000 iterates which we used for inference and model checking. By using the trace plots we “burn in” 1,000 iterates, and to nullify the effect of autocorrelations, we picked every tenth iterate thereafter. This rule was obtained by trial and error while tuning the Metropolis steps. We maintain the jumping probabilities in (0.25, 0.50); see Gelman *et al.* (1996).

3.3 Model Selection and Model Assessment

We used the minimum posterior predictive loss approach (Gelfand and Ghosh 1998) to select the best model among the first four.

Under squared error loss the minimum posterior predictive loss is

$$D_k = P + \frac{k}{k+1} G$$

$$P = \sum_{ij} \text{Var}(x_{ij}^{\text{pre}} | \mathbf{x}^{\text{obs}}), \quad G = \sum_{ij} \{E(x_{ij}^{\text{pre}} | \mathbf{x}^{\text{obs}}) - x_{ij}^{\text{obs}}\}^2$$

where $f(x_{ij}^{\text{pre}} | \mathbf{x}^{\text{obs}}) = \int f(x_{ij}^{\text{pre}} | \Omega) \pi(\Omega | \mathbf{x}^{\text{obs}}) d\Omega$ and x_{ij}^{pre} are the predicted values and Ω is the set of all parameters. This measure extends the one obtained earlier (Laud and Ibrahim 1995), and we have taken $k=100$ to match this earlier version. Note that for the nonresponse application, these measures are computed only on the complete BMI data after fitting our nonresponse models.

In Table 1 we present the deviance measure (D_{100}) and its associated components, goodness of fit (G) and the penalty (P) for the four models. Using the deviance measure the selection model is much better. While P is roughly the same, G is much smaller, making D_{100} smaller for the selection model. The difference between the two pattern mixture models are more pronounced than the difference between the two selection models. However, because standard errors are not available, it is difficult to tell the strength of the difference.

Table 1

Comparison of the Ignorable, Pattern Mixture and the Selection Models Using the Deviance Measure

Model	G	P	D_{100}
SEI	135	135	270
SE	118	135	253
PMI	268	135	403
PM	204	135	339

Note: $D_{100} = G + (100/(100+1)) P$ where G is a goodness of fit, P a penalty and D the deviance; the pattern mixture (PM) model and the selection model (SE) are both nonignorable. SEI is ignorable version of the selection model, PMI is ignorable version of the pattern mixture model.

Next, we look for deficiencies in the selection model. We use a Bayesian cross-validation analysis to assess the goodness of fit of the selected model (*i.e.*, the selection model). We do so by using deleted residuals on the respondents' BMI values.

Let $(\mathbf{x}_{(ij)}, \mathbf{r}_{(ij)})$ denote the vector of all observations excluding the $(i, j)^{\text{th}}$ observation (x_{ij}, r_{ij}) . Then, the $(i, j)^{\text{th}}$ deleted residual is given by

$$\text{DRES}_{ij} = \{x_{ij} - E(x_{ij} | \mathbf{x}_{(ij)}, \mathbf{r}_{(ij)})\} / \text{STD}(x_{ij} | \mathbf{x}_{(ij)}, \mathbf{r}_{(ij)}).$$

These values are obtained by performing a weighted importance sampling on the Metropolis-Hastings output. The posterior moments are obtained from

$$f(x_{ij} | \mathbf{x}_{(ij)}, \mathbf{r}_{(ij)}) = \int f(x_{ij} | \Omega) \pi(\Omega | \mathbf{x}_{(ij)}, \mathbf{r}_{(ij)}) d\Omega.$$

For the pattern mixture model

$$f(x_{ij} | \Omega) = f(x_{ij} | r_{ij} = 0, \Omega) p(r_{ij} = 0 | \Omega) + f(x_{ij} | r_{ij} = 1, \Omega) p(r_{ij} = 1 | \Omega)$$

and for the selection model

$$f(x_{ij} | \Omega) \sim \text{Normal} \{(\mathbf{z}'_{ij} \alpha_1 + v_{0i}) + (\mathbf{z}'_{ij} \alpha_2 + v_{1i}) a_{ij}, \sigma_3^2\}.$$

We also considered using the conditional posterior ordinate (CPO) which is $f(x_{ij} | \mathbf{x}_{(ij)}, \mathbf{r}_{(ij)})$ evaluated at the observed x_{ij} . However, these CPO's lead to similar results for identifying extremes.

We drew box plots (not shown) of DRES versus the four levels of race-sex and the thirty five counties, and they showed that the selection model fits well. We drew box plots of DRES versus age and, interestingly, we found a pattern. Age class 2-4 seems to fit well; the predicted BMI values are somewhat high for age class 5-9; and age classes 10-14 and 15-19 have larger variability. We look at the box plots of DRES versus age even further by separating out the box plots for 18 (*i.e.*, 2-19 years old) individual ages (see Figure 1). Ages 11-19 fits well, but there is a problem with ages 2-10 (*i.e.*, a downward curvature in the medians). The other three models show similar patterns. A further refinement of the selection model in section 5 fixes this problem.

4. Estimation and Prediction

In this section we perform an analysis on the NHANES III BMI data for children and adolescents (*i.e.*, 2-19 years old). We use the selection model, and then as a means to study sensitivity, we compare prediction under the non-ignorable nonresponse selection model with that of the other three models.

4.1 Estimation

We have studied the relation between BMI and age using 95% credible intervals for the parameters in the selection model. First, the interaction of race and sex is not important, but as expected there is an important relation of BMI on age. BMI increases substantially with age (95% credible interval for α_{21} is (11.89, 13.67)). The rate of increase for white males is smaller (95% credible interval for α_{22} is (-2.30, -0.19) and the 95% credible interval for α_{23} is (-3.03, -0.64)). Thus, while BMI increases with age, there is relatively less increase for white males. Apart from

the parameter θ_1 , which indicates strong nonignorability, the other parameters are essentially unimportant. For example, the 95% credible intervals for p_1 and p_2 are $(-0.53, 0.39)$ and $(-0.45, 0.45)$ respectively indicating that a simpler model can be used (*i.e.*, $p_1 = p_2 = 0$).

We take up the issue of ignorability further. We drew box plots (not shown) of the posterior densities of the β_{1i} , obtained from the iterates from the Metropolis-Hastings sampler, by county. All the box plots are above zero. This suggests that the nonresponse mechanism for each county is nonignorable. In addition, there are varying degrees of nonignorability. For example, several counties have the medians of the box plots near 1.5 while others have them near 2.

4.2 Prediction

It is desirable to predict the finite population mean BMI value and the proportion of respondents in the finite population. The sampled nonrespondents' BMI values are obtained through their conditional posterior densities included in the Metropolis-Hastings sampler. The non-sampled BMI values are to be predicted.

It is worthwhile noting that our models are applied to the logarithm of BMI with each individual having her/his covariates, and so the logarithm of each individual non-sampled value has to be predicted and then retransformed to

the original scale. However, the computation is reduced considerably because age, race and sex for each nonsampled individual is not known, but the number of individuals in each age-race-sex domain is known in the U.S. population by county.

The distributions of the nonsampled individuals are

$$f(x_{ij}, r_{ij} | \mathbf{x}^{\text{obs}}, \mathbf{r}^{\text{obs}}) = \int f(x_{ij}, r_{ij} | \Omega) \pi(\Omega | \mathbf{x}^{\text{obs}}, \mathbf{r}^{\text{obs}}) d\Omega,$$

$i = 1, \dots, \ell$, $j = n_i + 1, \dots, N_i$. For the pattern mixture model we have

$$f(x_{ij}, r_{ij} | \Omega) = f(x_{ij} | r_{ij}, \Omega) p(r_{ij} | \Omega)$$

and for the selection model we have

$$f(x_{ij}, r_{ij} | \Omega) = p(r_{ij} | x_{ij}, \Omega) f(x_{ij} | \Omega),$$

where Ω denote the set of all parameters.

Therefore, if we take a sample of size M from the posterior distribution, $\{\Omega^{(h)} : h = 1, \dots, M\}$, an estimator for $f(x_{ij}, r_{ij} | \mathbf{x}^{\text{obs}})$

$$f(x_{ij}, \hat{r}_{ij} | \mathbf{x}^{\text{obs}}) = M^{-1} \sum_{h=1}^M f(x_{ij}, r_{ij} | \Omega^{(h)}).$$

Thus, we can fill in the x_{ij} and r_{ij} for each $\Omega^{(h)}$ obtained from the MCMC algorithm from which we get M realizations $\bar{X}_i^{(h)}, P_i^{(h)}$, $h = 1, \dots, M$. Inference can now be made about \bar{X}_i in (1) and P_i in (2).

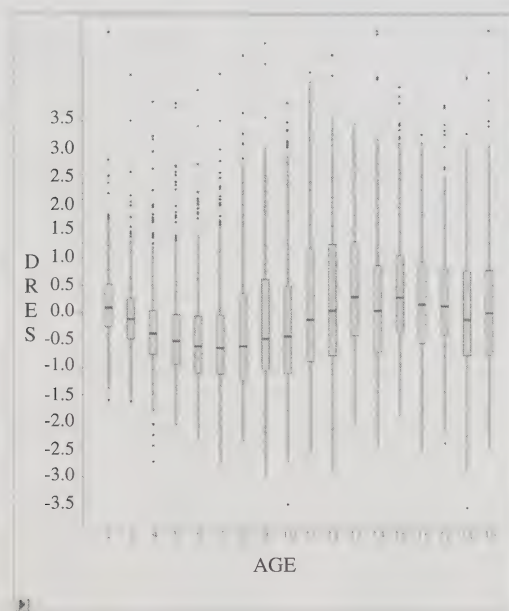


Figure 1. Box plots of the cross-validation residuals (DRES) by age for the selection model

We present 95% credible intervals for the finite population mean (FPM) BMI value and the finite population proportion (FPP) responding in order to judge sensitivity to the four models. Note that we provide these intervals for each domain: race by sex for each age class by county, and because they are very similar across domains we have presented in Table 2 the average of the end points of the credible intervals over county for black females only. The intervals for the FPM across the models are very similar. However, those for the FPP are very different. The intervals for the pattern mixture model and its ignorable version are similar except for age class 2–4. This is expected because these models express a linear regression of the logarithm of the odds of responding on age. The intervals for the FPP under the two pattern mixture models are essentially the same because they have the same relation with age, race, sex and their interaction. The intervals for the ignorable version of the selection model are all the same over age because in the response part of this model both age and BMI are ignored. We note that the intervals for the selection model have forms similar to the pattern mixture model and its ignorable version. As the intervals indicate, the FPM and FPP increase with age.

5. A Spline Regression Model

We now address the issue associated with the box plot in Figure 1. We have a further look at the observed data. A box plot of observed BMI values versus age shows that BMI is roughly constant for ages 2–8, then rises roughly linearly for ages 8–13, and finally rises very slowly for ages 14–19. This apparently important feature is not included in the four models. Thus, in this section we attempt to exploit this feature using a spline regression model.

We have used Part 1 of the selection model, and for Part 2 we use a join-point regression model. Generically, letting $c^+ = 0$ if $c \leq 0$ and $c^+ = c$ if $c > 0$, we take

$$x_{ij} = \varphi_{0ij} + \varphi_{1ij}(a_{ij} - 8)^+ + \varphi_{2ij}a_{ij} - 13^+ + e_{ij} \tag{9}$$

where in the spirit of our four models

$$\varphi_{kij} = z_{ij} \alpha_k + v_{ki}, \quad k = 0, 1, 2.$$

In (9) we have taken

$$e_{ij} \mid \sigma_3^2 \stackrel{\text{idd}}{\sim} \text{Normal}(0, \sigma_3^2)$$

and motivated by our earlier result (the v_{ki} are uncorrelated), rather than a trivariate normal density on $\mathbf{v}_i = (v_{1i}, v_{2i}, v_{3i})'$, we have taken

$$v_{ki} \mid \sigma_k^2 \stackrel{\text{idd}}{\sim} \text{Normal}(0, \sigma_k^2), \quad k = 0, 1, 2.$$

The distribution assumptions on the hyper-parameters remain unchanged.

We have computed the deviance measure for the spline model; see Table 1 for the other four models. For this model $G \approx 129$ and $P \approx 107$ compared with $G \approx 118$ and $P \approx 135$ for the selection model. That is, $D_{100} \approx 236$ for the spline regression model and $D_{100} \approx 253$ for the selection model. Thus, the spline regression model shows an improvement over the original selection model.

In Figure 2 we present box plots of DRES versus age. This is a much improved plot over the one for the selection model (see Figure 1). Observe that the medians fluctuate about 0 with very little variation. The box plots for ages 2, 3, 4, 5, 6 and 7 are a little less variable than the others. We also fit the quadratic join-point model in which we replace (9) by

$$x_{ij} = \varphi_{0ij} + \varphi_{1ij}(a_{ij} - 8)^+ + \varphi_{2ij}\{(a_{ij} - 13)^+\}^2 + e_{ij}$$

with all other assumptions remaining unchanged. This model did not show any substantial improvement over the alternative model specified by (9), which we retain without further refinement.

Table 2
Comparison of the Four Models Based on the Average Over All Counties of the End Points of the 95% Credible Intervals for the Finite Population Mean BMI (FPM) and Proportion (FPP) Responding for Black Females

Model		age			
		2–4	5–9	10–14	15–19
SEI	FPM	(14.80, 16.07)	(17.09, 18.58)	(19.63, 21.61)	(22.40, 25.19)
	FPP	(0.73, 0.79)	(0.73, 0.79)	(0.73, 0.79)	(0.73, 0.79)
SE	FPM	(15.55, 16.21)	(17.49, 18.36)	(19.52, 20.92)	(21.74, 23.91)
	FPP	(0.66, 0.78)	(0.71, 0.81)	(0.75, 0.84)	(0.78, 0.87)
PMI	FPM	(14.75, 16.10)	(17.04, 18.59)	(19.59, 21.55)	(22.42, 25.09)
	FPP	(0.49, 0.70)	(0.72, 0.84)	(0.84, 0.94)	(0.90, 0.98)
PM	FPM	(14.96, 15.79)	(17.16, 18.38)	(19.61, 21.45)	(22.37, 25.07)
	FPP	(0.49, 0.70)	(0.73, 0.84)	(0.84, 0.94)	(0.90, 0.98)

Note: SEI is ignorable version of the selection model, PMI is ignorable version of the pattern mixture model, PM is pattern mixture model, and SE is selection model.

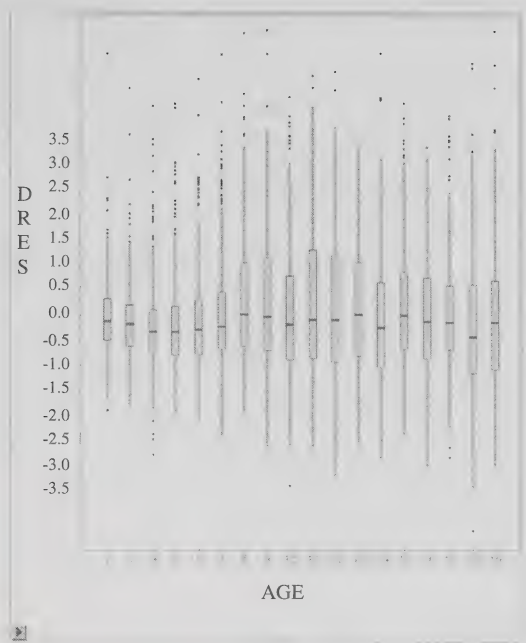


Figure 2. Box plots of the cross-validation residuals (DRES) by age for the spline regression model

In Table 3 we compare the FPM for the selection models (regression without splines and regression with splines). Again we average the end points of the 95% credible intervals over all counties. The intervals overlap suggesting similarity between the model without splines and the one with them. However, there are some exceptions. The largest difference between the intervals occur for individuals age 15–19 years old. In general, the spline model provides higher precision. For example, for age 10–19 the intervals for the spline model are contained by those for the model without the splines.

6. Conclusions

To analyze BMI data from NHANES III by age, race and sex within each county, (a) we have extended the normal-logistic regression model to a hierarchical Bayesian selection model, and (b) constructed a pattern mixture model and two ignorable nonresponse models to assess sensitivity to inference. A deviance measure shows that among the four models, the selection model is the best, and a cross-validation analysis shows that these models fit roughly equally well.

Table 3

Comparison of the Two Selection Models (Regression Without Splines and Regression with Splines) using the Average Over all Counties of the End Points of the 95% Credible Intervals for the Finite Population Mean BMI by Age, Race and Sex

R–S		age			
		2–4	5–9	10–14	15–19
BF	No Spline	(16.26, 16.92)	(16.44, 17.10)	(19.62, 21.41)	(21.35, 25.62)
	Spline	(15.65, 16.31)	(17.62, 18.41)	(19.70, 20.91)	(21.95, 23.82)
BM	No Spline	(16.10, 16.76)	(16.26, 16.92)	(18.83, 20.55)	(20.45, 24.53)
	Spline	(15.68, 16.32)	(17.32, 18.11)	(19.03, 20.21)	(20.84, 22.61)
OF	No Spline	(16.39, 17.00)	(16.56, 17.17)	(19.48, 21.19)	(21.16, 25.39)
	Spline	(16.01, 16.60)	(17.77, 18.54)	(19.62, 20.79)	(21.61, 23.38)
OM	No Spline	(16.53, 17.14)	(16.67, 17.29)	(19.22, 20.95)	(20.83, 24.98)
	Spline	(16.16, 16.74)	(17.74, 18.51)	(19.38, 20.55)	(21.13, 22.87)

Note: R–S is race-sex: BF is black female; BM is black male; OF is non-black female; and OM is non-black male.

Another contribution is the identification of a common deficiency in the selection model, the pattern mixture model and the two ignorable models. Based on the observed data, we have found that there is a dynamic relationship of BMI with age. Thus, we have further extended the selection model to include three linear splines. The cross validation analysis shows that there is an improvement over the selection model, and in fact, the deviance measure shows that the linear spline regression model is the best among the five models.

Our study on obesity is one of the key contributions in this work. The linear spline regression of BMI on age adjusting for race and sex, gives a better fit and improved precision than the selection model without splines. It is not easy to construct a model that is satisfactory for all aspects of the NHANES III data simultaneously. We have been able to do so for children and adolescents. BMI increases substantially with age; race and sex contributing negatively to this increase; there is relatively less increase for white males. In general, the effects of race and sex are relatively minor. There is some variation across the thirty five counties.

Appendix A The Pattern Mixture Model

For Part 1 of the pattern mixture model the response depends on age, race and sex, and the interaction of race and sex through the logistic regression

$$r_{ij} | \beta_i \stackrel{\text{iid}}{\sim} \text{Bernoulli} \left\{ e^{\beta_{0i} + \beta_{1i}a_{ij} + \beta_{2i}z_{ij1} + \beta_{3i}z_{ij2} + \beta_{4i}z_{ij3}} / (1 + e^{\beta_{0i} + \beta_{1i}a_{ij} + \beta_{2i}z_{ij1} + \beta_{3i}z_{ij2} + \beta_{4i}z_{ij3}}) \right\} \quad (\text{A.1})$$

$i = 1, \dots, L$, $j = 1, \dots, N_i$. Now, letting, $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i}, \beta_{4i})'$, note that while the vector β_i has $p = 5$ components, the corresponding vector in (4) has two components. Analogous to (4) we take

$$\beta_i | \theta, \Delta \stackrel{\text{iid}}{\sim} \text{Normal}(\theta, \Delta), \quad (\text{A.2})$$

and for the prior distribution,

$$\theta \sim \text{Normal}(\theta^{(0)}, \Delta^{(0)})$$

$$\text{and } \Delta^{-1} \sim \text{Wishart}\{(v^{(0)}\Lambda^{(0)})^{-1}, v^{(0)}\}, v^{(0)} > p, \quad (\text{A.3})$$

where $\theta^{(0)}, \Delta^{(0)}, \Lambda^{(0)}$ and $v^{(0)}$ are to be specified. Part 2 of this model for BMI incorporates a dependence on the response indicators, letting $w_{ij0} = 1$, $w_{ij1} = a_{ij}$,

$$x_{ij} = \sum_{t=0}^l (z'_{ij} \alpha_t + r_{ij} v_{it}) w_{ijt} + e_{ij}, \quad r_{ij} = 0, 1, \\ e_{ij} | \sigma_3^2 \stackrel{\text{iid}}{\sim} \text{Normal}(0, \sigma_3^2). \quad (\text{A.4})$$

The distributions on the (v_{0i}, v_{1i}) are the same as in (7). The prior distributions are exactly those in Part 2 of the selection model (i.e., see (6) and (7)).

We take $v^{(0)} = 2p$, a value that indicates near vagueness, maintains propriety and permits stability in computation. We show how to specify parameters like $\theta^{(0)}, \Delta^{(0)}, \alpha_t^{(0)}, \Delta_t^{(0)}, t = 1, 2, 3, \Lambda^{(0)}$ in Appendix C.

Appendix B Metropolis-Hastings Algorithm for Fitting the Selection Model

For the nonignorable nonresponse selection model the joint posterior density is

$$p(x^{(s, nr)}, \sigma^2, \alpha, \beta, v, \theta, \rho_1, \rho_2 | x^{(s, r)}) \propto \\ \prod_{i=1}^l \left\{ \prod_{j=1}^{r_i} \frac{1}{\sigma_3} e^{-\frac{1}{2\sigma_3^2} (x_{ij} - [z'_{ij}(\alpha_1 + a_{ij}\alpha_2) + v_{0i} + v_{1i}a_{ij}])^2} \frac{e^{\beta_{0i} + \beta_{1i}x_{ij}}}{1 + e^{\beta_{0i} + \beta_{1i}x_{ij}}} \right\} \\ \times \prod_{i=1}^l \left\{ \prod_{j=r_i+1}^{n_i} \frac{1}{\sigma_3} e^{-\frac{1}{2\sigma_3^2} (x_{ij} - [z'_{ij}(\alpha_1 + a_{ij}\alpha_2) + v_{0i} + v_{1i}a_{ij}])^2} \frac{1}{1 + e^{\beta_{0i} + \beta_{1i}x_{ij}}} \right\} \\ \times \left\{ \prod_{i=1}^l \frac{1}{\sigma_1 \sigma_2 \sqrt{1 - \rho_1^2}} \left[e^{-\frac{1}{2(1-\rho_1^2)} \left(\left(\frac{\beta_{0i} - \theta_0}{\sigma_1} \right)^2 - 2\rho_1 \left(\frac{\beta_{0i} - \theta_0}{\sigma_1} \right) \left(\frac{\beta_{1i} - \theta_1}{\sigma_2} \right) + \left(\frac{\beta_{1i} - \theta_1}{\sigma_2} \right)^2 \right)} \right] \right\} \\ \times \left\{ \prod_{i=1}^l \frac{1}{\sigma_4 \sigma_5 \sqrt{1 - \rho_2^2}} e^{-\frac{1}{2(1-\rho_2^2)} \left[\left(\frac{v_{0i}}{\sigma_4} \right)^2 - 2\rho_2 \left(\frac{v_{0i}}{\sigma_4} \right) \left(\frac{v_{1i}}{\sigma_5} \right) + \left(\frac{v_{1i}}{\sigma_5} \right)^2 \right]} \right\} \\ \times \left\{ \prod_{k=1}^5 \left(\frac{1}{\sigma_k^2} \right)^{\frac{a}{2} + 1} e^{-\frac{a}{2\sigma_k^2}} \left\{ e^{-\frac{1}{2}(\theta - \theta^{(0)})' \Delta^{(0)-1} (\theta - \theta^{(0)})} \right\} \right\} \\ \times \left\{ \prod_{k=1}^2 e^{-\frac{1}{2}(\alpha_k - \alpha_k^{(0)})' \Delta_k^{(0)-1} (\alpha_k - \alpha_k^{(0)})} \right\}. \quad (\text{B.1})$$

Let Ω denote the set of parameters $\beta, \theta, v, \alpha, \sigma_3^2, \psi_1, \psi_2$ and $x^{(s, nr)}$ where $\psi_1 = (\sigma_1^2, \sigma_2^2, \rho_1)'$ and $\psi_2 = (\sigma_4^2, \sigma_5^2, \rho_2)'$. Generically, let Ω_a denote all parameters in Ω except a ; for example, $\Omega_\beta = (\theta, v, \alpha, \sigma_3^2, \psi_1, \psi_2, x^{(s, nr)})$, so that the conditional posterior density (CPD) of β is denoted by $p(\beta | \Omega_\beta, x^{(s, r)})$. To perform the Metropolis-Hastings algorithm, one needs the CPD for each parameter given the others and $x^{(s, r)}$. Here we give a sketch of the algorithm.

The CPD for each of the parameters θ, v, α and σ_3^2 is easy to write down. But we need Metropolis steps for the CPD's of β, ψ_1, ψ_2 , and $x^{(s, nr)}$.

Conditioning on Ω_p , the parameters β_1, \dots, β_l , are independent with

$$p(\beta_i | \mathbf{x}^{(s,r)}) \propto \prod_{j=1}^{n_i} \left\{ \frac{e^{(\beta_{0i} + \beta_{1i} x_{ij}) r_{ij}}}{1 + e^{(\beta_{0i} + \beta_{1i} x_{ij})}} \right\} \times e^{-\frac{1}{2}(\beta_i - \theta_i)' \Delta_i^{-1} (\beta_i - \theta_i)},$$

where

$$\Delta_i = \begin{pmatrix} \sigma_1^2 & \rho_1 \sigma_1 \sigma_2 \\ \rho_1 \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}$$

and $x_{ij}, i=1, \dots, l$ and $j=r_i+1, \dots, n_i$ are to be predicted; see below. We use a technique based on logistic regression to obtain a multivariate Student's t proposal density in which tuning is obtained by varying its degree of freedom.

The method to draw from the CPD's of $\psi_1 = (\sigma_1^2, \sigma_2^2, \rho_1)$ and $\psi_2 = (\sigma_3^2, \sigma_4^2, \rho_2)$ is the same. The CPD of ψ_2 is

$$p(\psi_2 | \Omega_{\psi_2}, \mathbf{x}^{(s,r)}) \propto \left(\frac{1}{\sigma_4^2 \sigma_5^2} \right)^{\frac{a+l}{2}+1} e^{-\frac{b}{2} \left(\frac{1}{\sigma_4^2} + \frac{1}{\sigma_5^2} \right)} \times \frac{1}{(1-\rho_2)^{1/2}} e^{-\frac{1}{2(1-\rho_2^2)} \left\{ \frac{1}{\sigma_3^2} \sum_{i=1}^l v_{0i}^2 - \frac{2\rho_2}{\sigma_3 \sigma_4} \sum_{i=1}^l v_{0i} v_{1i} + \frac{1}{\sigma_4^2} \sum_{i=1}^l v_{1i}^2 \right\}}.$$

We have used the Fisher's z transformation (see Ruben 1966) to obtain a proposal density associated with normal distribution for $\log\{\rho_2/(1-\rho_2)\}$ and gamma distributions for σ_4^2 and σ_5^2 .

Finally, we consider the Metropolis step for drawing $\mathbf{x}^{(s, nr)} | \Omega_{\mathbf{x}^{(s, nr)}}, \mathbf{x}^{(s,r)}$. We note that in this CPD, $x_{ij}, i=1, \dots, l, j=r_i+1, \dots, n_i$, are independent with

$$p(x_{ij} | \Omega_{ij}, \mathbf{x}^{(s,r)}) \propto e^{-\frac{1}{2\sigma_3^2} [x_{ij} - \{z_{ij}(\alpha_1 + \alpha_2 a_{ij}) + v_{0i} + v_{1i} a_{ij}\}]^2} \left\{ 1 + e^{\beta_{0i} + \beta_{1i} x_{ij}} \right\}^{-1}.$$

We have constructed a proposal density using least squares techniques. We note that the proposal density Normal($z_{ij}(\alpha_1 + \alpha_2 a_{ij}) + v_{0i} + v_{1i} a_{ij}, \sigma_3^2$) did not perform well (see Chib and Greenberg 1995).

Appendix C Specification of Hyperparameters

We discuss how to specify the hyperparameters $(\theta^{(0)}, \Delta^{(0)})$ and $(\alpha_k^{(0)}, \Gamma_k^{(0)})$, $k=1, 2$, associated with θ and α_k , $k=1, 2$ in the selection model.

First, consider $(\theta^{(0)}, \Delta^{(0)})$. For $i=1, \dots, l, j=1, \dots, n_i$ fit the logistic regression model $r_{ij} \stackrel{\text{iid}}{\sim} \text{Bernoulli}\{e^{\beta_{0i} + \beta_{1i} x_{ij}} / (1 + e^{\beta_{0i} + \beta_{1i} x_{ij}})\}^{-1}$, where x_{ij} are obtained by prediction (see Appendix A). Letting $\hat{\beta}_i, i=1, \dots, l$ denote

the least squares estimators, we assume that $\hat{\beta}_i \stackrel{\text{iid}}{\sim} \text{Normal}(\theta^{(0)}, \tilde{\Delta}^{(0)})$ to get $\theta^{(0)} = 1/l \sum_{i=1}^l \hat{\beta}_i$ and

$$\tilde{\Delta}^{(0)} = \frac{1}{l-1} \sum_{i=1}^l (\hat{\beta}_i - \theta_{(0)}) (\hat{\beta}_i - \theta_{(0)})' \quad (\text{C.1})$$

and we set $\Delta^{(0)} = \kappa_1 \tilde{\Delta}^{(0)}$, where κ_1 is to be selected.

Next, we consider how to specify $(\alpha_k^{(0)}, \Gamma_k^{(0)})$, $k=1, 2$. We fit $x_{ij} = z_{ij}'(\alpha_1 + \alpha_2 a_{ij}) + e_{ij}$, where a_{ij} is the age of the j^{th} individual in the i^{th} county, $i=1, \dots, l, j=1, \dots, n_i$ to get least squares estimators, $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2)$ and its covariance matrix $\hat{\Gamma}^{(0)}$. We set $\alpha_k^{(0)} = \hat{\alpha}_k$, and $\Gamma_k^{(0)} = \kappa_2 \hat{\Gamma}_k^{(0)}$, where $\hat{\Gamma}_k^{(0)}, k=1, 2$ is the corresponding block matrix of $\hat{\Gamma}^{(0)}, k=1, 2$ and κ_2 is to be specified.

We have experimented with κ_1 in (C.1). We used $\kappa_1=100$ to provide a proper diffuse prior; a value of $\kappa_1=1,000$ did not change our predictions. Similarly, we used $\kappa_2=100$.

References

- Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.
- Cohen, G., and Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.
- Cowles, M., and Carlin, B. (1996). Markov chain Monte Carlo diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.
- Daniels, M.J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27, 569-580.
- David, M., Little, R.J.A., Samuel, M.E. and Triest, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.
- Dietz, W.H. (1998). Health consequences of obesity in youth: Childhood predictors of adult disease. *Pediatrics*, 101, 518-525.
- Fagot-Campagna, A. (2000). Emergence of type 2 diabetes mellitus in children: Epidemiological evidence. *Journal of Pediatric Endocrinology Metabolism*, 13, 1395-1405.
- Gelfand, A., and Ghosh, S. (1998). Model choice: A minimum posterior predictive approach. *Biometrika*, 85, 1-11.
- Gelfand, A., Sahu, S. and Carlin, B. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82, 479-488.
- Gelman, A., Roberts, G.O. and Gilks, W.R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics* (Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford, U.K.: Oxford University Press, 599-607.
- Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475-492.

- Kuczmarski, R.J., Carroll, M.D., Flegal, K.M. and Troiano, R.P. (1997). Varying body mass index cutoff points to describe overweight prevalence among U.S. adults: NHANES III (1988 to 1994). *Obesity Research*, 5, 542-548.
- Laud, P., and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B*, 57, 247-262.
- Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.
- Mohadjer, L., Bell, B. and Waksberg, J. (1994). National health and Nutrition Examination Survey III-Accounting for item nonresponse bias. Internal Report, National Center for Health Statistics.
- Nandram, B., and Choi, J.W. (2002 a). A hierarchical Bayesian nonresponse model for binary data with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Nandram, B., and Choi, J.W. (2002 b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.
- Nandram, B., Han, G. and Choi, J.W. (2002). A hierarchical bayesian nonignorable nonresponse model for multinomial data from small areas. *Survey Methodology*, 28, 145-156.
- Natarajan, R., and Kass, R.E. (2000). Reference Bayesian methods for generalized linear models. *Journal of the American Statistical Association*, 95, 227-237.
- National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics Series 2*, 113.
- National Center for Health Statistics (1994). Plan and operation of the third national health and nutrition examination survey. *Vital and Health Statistics Series*, 1, 32.
- Ogden, C.L., Flegal, K.M., Carroll, M.D. and Johnson, C.L. (2002). Prevalence and trends in overweight among us children and adolescents, 1999-2000. *Journal of the American Medical Association*, 288, 1728-1732.
- Olson, R.L. (1980). A least square correction for selectivity bias. *Econometrica*, 48, 1815-1820.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons, Inc.
- Ruben, H. (1966). Some new results on the distribution of the sample correlation coefficient. *Journal of the Royal Statistical society, Series B*, 28, 513-525.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.
- Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.
- Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M. Little, R.J. A. and Rubin, D.B. (1996). The NHANES III multiple imputation project. *Survey research methods, Proceedings of the American Statistical Association*, 28-37.
- Squires, S. (2001). National plan urges to combat obesity: Weight-related illnesses kill 300,000 Americans annually, Surgeon General says. *The Washington Post*, December 14, 2001.
- Stasny, E.A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the National Crime Survey. *Journal of the American Statistical Association*, 86, 296-303.
- Who Consultation on Obesity (2000). Obesity: Preventing and managing the global epidemic. *WHO Technical Report Series 894*, Geneva, Switzerland: World Health Organization.
- Wright, C.M., Parker, L., Lamont, D. and Craft, A.W. (2001). Implications of childhood obesity for adult health: Findings from thousand families cohort study. *British Medical Journal*, 323, 1280-1284.

Towards Nonnegative Regression Weights for Survey Samples

Mingue Park and Wayne A. Fuller¹

Abstract

Procedures for constructing vectors of nonnegative regression weights are considered. A vector of regression weights in which initial weights are the inverse of the approximate conditional inclusion probabilities is introduced. Through a simulation study, the weighted regression weights, quadratic programming weights, raking ratio weights, weights from logit procedure, and weights of a likelihood-type are compared.

Key Words: Raking ratio; Maximum likelihood; Quadratic programming; Simple Conditionally Weighted (SCW) estimator.

1. Introduction

In survey sampling, information about the population is often available at the analysis stage. One method of using this information is through regression estimation. There are a number of ways to construct a regression estimator of the population mean or total. One regression estimator of the mean is

$$\bar{y}_{\text{reg}} = \sum_{i=1}^n w_i y_i = \bar{y}_{\pi} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\pi})' \tilde{\boldsymbol{\beta}}, \quad (1)$$

where

$$w_i = \alpha_i + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\pi})' \left(\sum_{j=1}^n \bar{\mathbf{x}}_j' \phi_{jj}^{-1} \mathbf{x}_j \right)^{-1} \mathbf{x}_i \phi_{ii}^{-1}, \quad (2)$$

$$(\bar{y}_{\pi}, \bar{\mathbf{x}}_{\pi}) = \left(\sum_{i=1}^n \pi_i^{-1} \right)^{-1} \sum_{i=1}^n \pi_i^{-1} (y_i, \mathbf{x}_i) =: \sum_{i=1}^n \alpha_i (y_i, \mathbf{x}_i),$$

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \mathbf{x}_i' \phi_{ii}^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^n \mathbf{x}_i' \phi_{ii}^{-1} y_i,$$

$$\alpha_i = \left(\sum_{j=1}^n \pi_j^{-1} \right)^{-1} \pi_i^{-1},$$

$\Phi = \text{diag}(\phi_{11}, \dots, \phi_{nn})$ is a nonsingular diagonal matrix, the π_i 's are the selection probabilities and $\bar{\mathbf{x}}_N$ is the population mean of \mathbf{x} . A possible choice of ϕ_{ii}^{-1} is α_i . A review of the use of such information in regression estimation for sample surveys is given by Fuller (2002).

It is well known that regression weights that are used to define a regression estimator such as (2) can be very large or (and) can be negative. If the regression weights are to be used to estimate a finite population total in a general purpose survey, it seems reasonable that no individual weight

should be less than one. Also, it seems reasonable, on robustness grounds, to avoid very large weights.

There are several ways to construct regression weights with a reduced range of values. Huang and Fuller (1978) defined a procedure to modify the w_i so that there are no negative weights and no large weights. Husain (1969) suggested quadratic programming as a procedure to place bounds on the weights. Quadratic programming and a number of other procedures build on the fact that the weights can be defined as values that optimize some function. Deville and Särndal (1992) considered seven objective functions that can be used to construct weights. They suggested objective functions that can be used to produce weights which fall within a given range. Deville, Särndal and Sautory (1993) introduced the program, CALMAR, written as a SAS macro that can be used to calculate weights corresponding to four different objective functions when auxiliary information in the survey consists of known marginal counts in a frequency table.

Another modification of regression weights is to relax some of the restrictions used in constructing the estimator. Husain (1969) considered modifying weights for a simple random sample from a normal distribution. He derived the weights that minimize the mean square error (MSE) of the resulting estimator. Bardsley and Chambers (1984) considered an estimator based on an objective function and the division of the auxiliary variable into two components. They studied the behavior of the estimator from a model perspective. Rao and Singh (1997) studied an estimator in which tolerances are given for the difference between the final estimator for part of the auxiliary variables vector and the corresponding elements of the population vector.

In this paper, we consider different types of regression weights including a procedure based on Tillé's (1998) conditional selection probabilities. The approximate conditional

1. Mingue Park, University of Nebraska, 103 Miller Hall, Lincoln, NE, 68588-0712, U.S.A.; Wayne A. Fuller, Iowa State University, 221 Snedecor Hall, Ames, IA 50011-1210, U.S.A.

inclusion probabilities are used to compute regression weights that are positive for most samples. These regression weights are compared to raking ratio weights, to quadratic programming weights, weights from logit procedure, and to weights based on a likelihood-type objective function.

2. Maximum Likelihood and Raking Ratio

Consider a two-way table with r rows and c columns. The population cell U_{ij} contains N_{ij} elements; $i=1, \dots, r, j=1, \dots, c$. Assume marginal counts $N_{i\cdot}, N_{\cdot j}$ are known. The population characteristics of interest are the N_{ij} or, equivalently, $p_{ij} = N^{-1} N_{ij}$. For a simple random nonreplacement sample of size n , Deming and Stephan (1940) suggested a raking ratio procedure to get the solution for the cell frequencies. See also Stephan (1942). If we assume the sample is a random sample from a multinomial distribution defined by the population entries in a two way table, we can construct an estimator using the maximum likelihood procedure.

Deville and Särndal (1992) defined a class of calibration estimators, \bar{y}_{cal} , of the population mean of y as

$$\bar{y}_{\text{cal}} = \sum_{i=1}^n w_i y_i, \quad (3)$$

where the w_i 's minimize the objective function $\sum_{i=1}^n G(w_i, \alpha_i)$ subject to constraints

$$\sum_{i=1}^n w_i = 1 \quad \text{and} \quad \sum_{i=1}^n w_i \mathbf{x}_i = \bar{\mathbf{x}}_N, \quad (4)$$

and $G(w_i, \alpha_i)$ is a measure of distance between an initial weight α_i and a final weight w_i . The raking ratio and maximum likelihood estimators of the population cell fraction, p_{ij} , belong to the class of calibration estimators.

The raking ratio weights for the population cell fraction, with a simple random sample, can be obtained by minimizing

$$\sum_{k=1}^n w_k \log \left(\frac{w_k}{n^{-1}} \right) - w_k + n^{-1}, \quad (5)$$

subject to the constraints (4) with

$$\mathbf{x}_k = (\delta_{1\cdot}, \dots, \delta_{r\cdot}, \delta_{\cdot 1}, \dots, \delta_{\cdot c}), \quad (6)$$

where $\delta_{i\cdot} = 1$ if k^{th} element belongs to the i^{th} row and $\delta_{i\cdot} = 0$ otherwise, and $\delta_{\cdot j} = 1$ if k^{th} element belongs to the j^{th} column and $\delta_{\cdot j} = 0$ otherwise. The raking ratio estimator for the population cell fraction p_{ij} is the estimator (3) where $y_k = 1$ if the k^{th} element belongs to cell ij and $y_k = 0$ otherwise.

For the maximum likelihood estimator of the population fraction, with a simple random sample, Deville and Särndal (1992) suggested minimizing

$$\sum_{k=1}^n -n^{-1} \log \left(\frac{w_k}{n^{-1}} \right) + w_k - n^{-1} \quad (7)$$

subject to (4) with \mathbf{x} defined in (6).

Chen and Sitter (1999) suggested a *pseudo empirical likelihood estimator*. They defined the population likelihood of y_i as

$$\sum_{i=1}^N \log w_{i,U}, \quad (8)$$

where $w_{i,U}$ is the density at observation y_i . With a sample of size n , they suggested the pseudo empirical likelihood estimator of the form

$$\bar{y}_{\text{EL}} = \sum_{i=1}^n w_i y_i, \quad (9)$$

where w_i 's are obtained by minimizing the function

$$-\sum_{i=1}^n \pi_i^{-1} \log w_i, \quad (10)$$

under the restrictions (4). The resulting w_i are equal to those obtained by minimizing (7) with $=N\pi_i$ under the restrictions (4).

Deville and Särndal (1992) showed that the raking ratio and maximum likelihood estimators are approximately equal to a regression estimator of the form (1), and, hence, have the same limiting distribution as the regression estimator. Weights for the raking ratio and maximum likelihood estimators are nonnegative if the solutions for the weights exist.

3. Weighted Regression Using Conditional Probabilities

Tillé (1998) suggested the use of approximate conditional inclusion probabilities, conditioning on the Horvitz-Thompson estimators of auxiliary variables, to compute an estimator for the population mean of the study variable. His approximation can be extended to produce regression weights that are nonnegative with high probability.

Assume that the vector of population means of auxiliary variables, $\bar{\mathbf{x}}_N$, is known. Consider the Horvitz-Thompson estimator of $\bar{\mathbf{x}}_N$ given by

$$\bar{\mathbf{x}}_{\text{HT}} = \frac{1}{N} \sum_{i=1}^n \frac{\mathbf{x}_i}{\pi_i}, \quad (11)$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ and π_i is the unconditional inclusion probability. Tillé (1998) introduced the simple conditionally weighted (SCW) estimator,

$$\bar{y}_{p\pi} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_{i|\bar{\mathbf{x}}_{\text{HT}}}}, \quad (12)$$

where $\pi_{i|\bar{\mathbf{x}}_{\text{HT}}}$ is the conditional inclusion probability of the i^{th} element conditioning on $\bar{\mathbf{x}}_{\text{HT}}$. To construct the SCW-estimator of \bar{y}_N , the conditional inclusion probability $\pi_{i|\bar{\mathbf{x}}_{\text{HT}}}$ is required. If $\bar{\mathbf{x}}_{\text{HT}}$ takes the value \mathbf{t} , we have

$$\pi_{i|\bar{\mathbf{x}}_{\text{HT}}} = \pi_i \frac{P\{\bar{\mathbf{x}}_{\text{HT}} = \mathbf{t} | i \in A\}}{P\{\bar{\mathbf{x}}_{\text{HT}} = \mathbf{t}\}}, \quad (13)$$

where A is the set of indices for the sample elements.

In order to compute the conditional inclusion probabilities, it is necessary to know the probability distribution of $\bar{\mathbf{x}}_{\text{HT}}$ unconditionally and conditionally on the presence of each unit in the sample. Except for some particular cases, this probability distribution is very complex. For this reason, approximation of the conditional inclusion probability is considered.

Under the assumption that $\bar{\mathbf{x}}_{\text{HT}}$ has an approximately normal distribution unconditionally and conditionally on the presence of each unit in the sample, the conditional inclusion probability (13) can be approximated by

$$\hat{\pi}_{i|\bar{\mathbf{x}}_{\text{HT}}} = \pi_i \left| \frac{\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}}}{\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}} \right|^{1/2} \exp\{0.5(\mathbf{G}_{\bar{\mathbf{x}}\bar{\mathbf{x}}} - \mathbf{G}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)})'\}, \quad (14)$$

where $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}} = \text{Var}\{\bar{\mathbf{x}}_{\text{HT}} | F\}$, $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)} = \text{Var}\{\bar{\mathbf{x}}_{\text{HT}} | F, i \in A\}$,

$$\mathbf{G}_{\bar{\mathbf{x}}\bar{\mathbf{x}}} = (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N) \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}}^{-1} (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N)',$$

$$\mathbf{G}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)} = (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_{N,(i)}) \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}^{-1} (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_{N,(i)})',$$

$$\bar{\mathbf{x}}_{N,(i)} = E\{\bar{\mathbf{x}}_{\text{HT}} | F, i \in A\} =$$

$$(N\pi_i)^{-1} \mathbf{x}_i + N^{-1} \sum_{\substack{j=1 \\ j \neq i}}^n (\pi_i \pi_j)^{-1} \pi_{ij} \mathbf{x}_j,$$

A is the set of indices appearing in the sample and $F = \{y_1, \dots, y_N\}$ is the finite population. Tillé (1998) gives an expression for $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}$ for the general case.

Assume the design covariance matrices $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}}$ and $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}$ are positive definite and assume the vector of auxiliary variables is normally distributed. Tillé (1999) showed that the SCW-estimator defined in (12) with the approximate conditional inclusion probabilities of (14) satisfies

$$\bar{y}_{p\pi} = \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\beta}_N + O_p(n^{-1}) \quad (15)$$

$$= \bar{y}_{\text{reg}} + O_p(n^{-1}), \quad (16)$$

where

$$\hat{\beta}_N = \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}}^{-1} \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{y}}},$$

$$\bar{y}_{\text{reg}} = \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\beta},$$

$$\hat{\beta} = \hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}}}^{-1} \hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{y}}} = (\mathbf{X}' \Phi^{-1} \mathbf{X})^{-1} \mathbf{X}' \Phi^{-1} \mathbf{y},$$

$\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_n)'$, $\mathbf{y} = (y_1, \dots, y_n)'$, the ij^{th} element of Φ^{-1} is $N^{-2}(\pi_{ij} \pi_i \pi_j)^{-1}(\pi_{ij} - \pi_i \pi_j)$, $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}}$ is the design variance of $\bar{\mathbf{x}}_{\text{HT}}$, $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{y}}}$ is the design covariance of $\bar{\mathbf{x}}_{\text{HT}}$ and \bar{y}_{HT} , $\hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}}}$ is the Horvitz-Thompson variance estimator of $\bar{\mathbf{x}}_{\text{HT}}$, and $\hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{y}}}$ is the Horvitz-Thompson estimator of the covariance of $\bar{\mathbf{x}}_{\text{HT}}$ and \bar{y}_{HT} .

Given a complex design, a number of the quantities in (14) are difficult to compute. However, approximations giving the same large sample properties for the estimator are relatively easy to compute. We replace $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}}$ and $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}$ with estimators, replace $\bar{\mathbf{x}}_{N,(i)}$ with $\bar{\mathbf{x}}_N + \mathbf{d}_{x_i}$, define

$$\hat{\mathbf{M}}_{\bar{\mathbf{x}}\bar{\mathbf{y}}} = \sum_{i \in A} (N\pi_i)^{-1} \mathbf{d}'_{x_i} y_i, \quad (17)$$

and assume

$$\text{Var}\{n(\hat{\mathbf{M}}_{\bar{\mathbf{x}}\bar{\mathbf{y}}} - \mathbf{M}_{\bar{\mathbf{x}}\bar{\mathbf{y}}})\} = O(n^{-1}), \quad (18)$$

$$\mathbf{d}_{x_i} = O_p(n^{-1}), \quad (19)$$

where \mathbf{d}_{x_i} is a function of the sample and $\mathbf{M}_{\bar{\mathbf{x}}\bar{\mathbf{y}}}$ is a population quantity. Often $\mathbf{M}_{\bar{\mathbf{x}}\bar{\mathbf{y}}}$ is the population covariance matrix $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{y}}}$, but this equality is not required in order for the estimator to be well defined. In many cases one can compute \mathbf{d}_{x_i} as a multiple of the jackknife deviate. Also in many situations, an adequate value for the estimator, $\hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}$, of $\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}$ is $n^{-1}(n-1)\hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}}}$. We write our generalization of (14) as

$$\tilde{\pi}_{i|\bar{\mathbf{x}}_{\text{HT}}} = \pi_i \left| \frac{\hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}}}}{\hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}} \right|^{1/2} \exp\{0.5(\hat{\mathbf{G}}_{\bar{\mathbf{x}}\bar{\mathbf{x}}} - \tilde{\mathbf{G}}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)})'\}, \quad (20)$$

where

$$\hat{\mathbf{G}}_{\bar{\mathbf{x}}\bar{\mathbf{x}}} = (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N) \hat{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}}}^{-1} (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N)',$$

$$\tilde{\mathbf{G}}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)} = (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N - \mathbf{d}_{x_i}) \tilde{\Sigma}_{\bar{\mathbf{x}}\bar{\mathbf{x}},(i)}^{-1} (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N - \mathbf{d}_{x_i})'.$$

Let the estimator (12) constructed with the $\tilde{\pi}_{i|\bar{\mathbf{x}}_{\text{HT}}}$ of (20) be

$$\bar{y}_{p\tilde{\pi}} = N^{-1} \sum_{i=1}^n \tilde{\pi}_{i|\bar{\mathbf{x}}_{\text{HT}}}^{-1} y_i. \quad (21)$$

An approximate conditional inclusion probability with a simple random sample and a single auxiliary variable is

$$\tilde{\pi}_{i|\bar{x}_n} = \frac{n}{N} \left[\frac{\hat{\sigma}_{\bar{x}}}{\hat{\sigma}_{\bar{x},(i)}} \right] \exp \left\{ \frac{1}{2} \left[\frac{(\bar{x}_n - \bar{x}_N)^2}{\hat{\sigma}_{\bar{x}}^2} - \frac{(\bar{x}_n - \bar{x}_N - d_{x_i})^2}{\hat{\sigma}_{\bar{x},(i)}^2} \right] \right\},$$

where

$$d_{x_i} = [n(N-1)]^{-1} (N-n) (x_i - \bar{x}_N),$$

$$\hat{\sigma}_{\bar{x},(i)}^2 = \frac{(N-n)(n-1)}{n^2(N-2)} \left[s_x^2 - \frac{N(x_i - \bar{x}_N)^2}{(N-1)^2} \right] \approx \frac{n-1}{n} \hat{\sigma}_{\bar{x}}^2,$$

$$\hat{\sigma}_{\bar{x}}^2 = (n^{-1} - N^{-1}) s_x^2,$$

and

$$s_x^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

In this case, $d_{x_i} = \bar{x}_{N,(i)} - \bar{x}_N$ and $M_{\bar{x}\bar{y}} = \text{Cov}(\bar{x}_{\text{HT}}, \bar{y}_{\text{HT}})$.

The SCW-estimator (21) with the approximate conditional inclusion probabilities is not calibrated, that is, the estimator (21) for the mean of the vector of auxiliary variables is not the vector of population means. It is relatively easy to standardize the probabilities so that they sum to one or sum to the stratum fraction for stratified sampling. To construct a calibrated estimator for the general case, we suggest computing the regression estimator with $[\sum_{j=1}^n \tilde{\pi}_{j|\bar{x}_{\text{HT}}}^{-1}]^{-1} \tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-1}$ as initial weights. The suggested estimator is

$$\begin{aligned} \bar{y}_{\text{wreg}} &= \bar{y}_c + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_c) \hat{\beta}_{c,1} \\ &= \sum_{i=1}^n w_i y_i, \end{aligned} \quad (22)$$

where

$$\begin{aligned} (\bar{y}_c, \bar{\mathbf{x}}_c) &= \sum_{i=1}^n \alpha_i (y_i, \mathbf{x}_i), \\ (\hat{\beta}_{c,0}, \hat{\beta}_{c,1})' &= \left[\sum_{i=1}^n \alpha_i \mathbf{z}_i' \mathbf{z}_i \right]^{-1} \left[\sum_{i=1}^n \alpha_i \mathbf{z}_i' y_i \right], \\ \mathbf{z}_i &= (1, \mathbf{x}_i - \bar{\mathbf{x}}_c), \\ \alpha_i &= \left[\sum_{j=1}^n \tilde{\pi}_{j|\bar{x}_{\text{HT}}}^{-1} \right]^{-1} \tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-1}, \end{aligned}$$

$$w_i = \alpha_i$$

$$+ (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_c) \left[\sum_{j=1}^n \alpha_j (\mathbf{x}_j - \bar{\mathbf{x}}_c)' (\mathbf{x}_j - \bar{\mathbf{x}}_c) \right]^{-1} \alpha_i (\mathbf{x}_i - \bar{\mathbf{x}}_c)',$$

and $\tilde{\pi}_{i|\bar{x}_{\text{HT}}}$ is the approximate conditional inclusion probability of (20). We assume the vector of auxiliary variables contains one so that the estimator is location invariant.

The estimator (21) is approximately equal to a regression estimator and estimator (22) is also approximately equal to the same regression estimator.

Theorem: Let a sequence of populations and samples, $\{F_N, A_N\}$, satisfy

$$(\bar{y}_{\text{HT}}, \bar{\mathbf{x}}_{\text{HT}}) - (\bar{y}_N, \bar{\mathbf{x}}_N) = O_p(n^{-1/2}). \quad (23)$$

Assume that the sequences of estimated covariance matrices, $\hat{\Sigma}_{\bar{x}\bar{x}}$ and $\hat{\Sigma}_{\bar{x}\bar{x},(i)}$, satisfy

$$\begin{aligned} [\mathbf{D}^{-1/2} \hat{\Sigma}_{\bar{x}\bar{x},(i)} \mathbf{D}^{-1/2}]^{-1} \\ - [\mathbf{D}^{-1/2} \hat{\Sigma}_{\bar{x}\bar{x}} \mathbf{D}^{-1/2}]^{-1} = O_p(n^{-1}), \end{aligned} \quad (24)$$

where \mathbf{D} denotes a diagonal matrix having the elements of the diagonal of $\hat{\Sigma}_{\bar{x}\bar{x}}$ on its diagonal. Let \mathbf{d}_{x_i} be a function of the sample satisfying (19) and assume (18) holds. Assume the sequence of Horvitz-Thompson variance estimators satisfies

$$\text{Var} \left\{ n \left[\text{Vech} \left(\hat{\Sigma}_{\bar{z}\bar{z}, \text{HT}} - \Sigma_{\bar{z}\bar{z}} \right) \right] \right\} = O(n^{-1}), \quad (25)$$

where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ and $\Sigma_{\bar{z}\bar{z}}$ is positive definite. Assume $E\{\tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-2}\}$ is bounded, where $\tilde{\pi}_{i|\bar{x}_{\text{HT}}}$ is defined in (20). Then, the SCW-estimator $\bar{y}_{p\bar{\pi}}$ of (21) satisfies

$$\begin{aligned} \bar{y}_{p\bar{\pi}} &= \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \boldsymbol{\theta}_N + O_p(n^{-1}) \\ &= \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\boldsymbol{\theta}} + O_p(n^{-1}), \end{aligned}$$

where $\hat{\boldsymbol{\theta}} = \hat{\Sigma}_{\bar{x}\bar{x}}^{-1} \hat{\mathbf{M}}_{\bar{x}\bar{y}}$ and $\boldsymbol{\theta}_N = \Sigma_{\bar{x}\bar{x}}^{-1} \mathbf{M}_{\bar{x}\bar{y}}$.

If $\text{Var}\{\sum_{i=1}^n \pi_i^{-1}\} > 0$, assume \mathbf{x}_i contains one as an element. Assume $\mathbf{M}_{\bar{x}\bar{y}} = \Sigma_{\bar{x}\bar{y}}$. Then the weighted regression estimator of (22) satisfies

$$\bar{y}_{\text{wreg}} = \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\boldsymbol{\theta}} + O_p(n^{-1}).$$

For proof, see the appendix.

To illustrate the nature of the different types of regression weights, we selected a simple random sample of size 40 from a normal population with mean zero and variance one. The sample mean is -0.614 and the population mean is zero. The weight for the regression estimator is given by (2) with $\alpha_i = \phi_{ii}^{-1} = n^{-1}$. The weights for the raking ratio and MLE are obtained by minimizing the objective functions (5) and (7), respectively, under the restriction (4). Weights for the SCW-weighted regression estimator are given in (22). The weights are plotted against the sample x values in Figure 1. Five of the simple regression weights are less than zero because of the large discrepancy between the sample and the population means. All weights for the SCW-weighted regression estimator, MLE and raking ratio are nonnegative. Figure 1 shows that the behaviors of raking ratio and SCW-weighted regression weights are similar and that MLE has an extremely large weight in this sample.

Table 1 contains selected weights for the smallest x values, x values close to the sample mean, x values close to the population mean, and the largest x values. For the x -values farthest from the population mean MLE gives the largest weights. For x -values near the sample mean the ordinary least squares weights are close to n^{-1} while the other weights are less than n^{-1} . The MLE weights are close to n^{-1} for x -values close to the population mean while the other weights are larger.

Table 1
Selected Regression Weights for Illustrated Example

x	Weights multiplied by $n = 40$			
	Reg	W. Reg	Raking	MLE
-2.103	-0.56	0.12	0.16	0.40
-1.941	-0.40	0.12	0.20	0.40
-1.727	-0.16	0.20	0.24	0.44
-0.710	0.88	0.68	0.68	0.68
-0.670	0.96	0.72	0.68	0.68
-0.468	1.16	0.88	0.84	0.76
-0.103	1.52	1.28	1.24	0.92
0.021	1.68	1.44	1.40	1.00
0.097	1.76	1.56	1.52	1.08
0.628	2.32	2.60	2.60	1.84
0.662	2.36	2.68	2.72	1.92
1.237	2.96	4.60	4.88	9.12

Simulation Study

To compare the alternative methods of constructing regression weights we conducted a simulation study. A total of 30,000 simple random samples of size 32 were selected from a χ^2 distribution with two degrees of freedom. The parameters being estimated are those of the infinite

generating mechanism. Let x_i be the value for the i^{th} sampled element. Six estimation procedures were considered.

- 1. Ordinary least squares regression (OLS)
- 2. Quadratic programming with upper and lower bounds (QP)
- 3. Weighted regression with SCW weights (SCW reg)
- 4. Maximum likelihood objective function (MLE)
- 5. Raking objective function (Raking reg)
- 6. Logit procedure with upper and lower bounds (Logit)

The weights for the OLS estimator were calculated by (2) with $\alpha_i = n^{-1}$. The quadratic programming weights minimize $\sum_{i=1}^n w_i^2$ subject to the constraint $0 \leq w_i \leq 0.065$ for all i and subject to constraints (4). The quadratic programming procedure is equivalent to the truncated linear method of case 7 of Deville and Särndal (1992). Weights for the SCW weighted regression were calculated by minimizing $\sum_{i=1}^n \alpha_i^{-1} w_i^2$ subject to constraints (4), where α_i is defined in (22). The weights for raking and maximum likelihood were obtained by minimizing the objective functions (5) and (7), respectively, under the restriction (4). Weights calculated by the logit procedure minimize the function $\sum_{i=1}^n G(nw_i)$ subject to constraints (4), where

$$G(nw_i) = a^{-1} \left[(nw_i) \ln(nw_i) + (u - nw_i) \ln \left(\frac{u - nw_i}{u - 1} \right) \right],$$

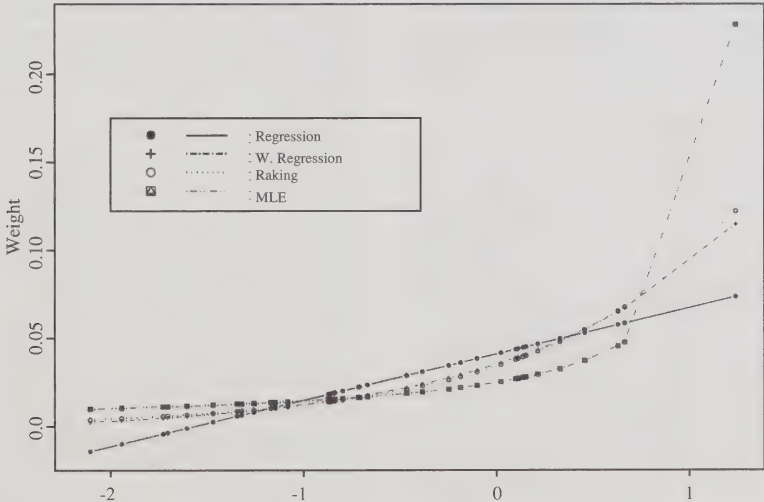


Figure 1. Comparison of four sets of weights.

if $0 < nw_i < u$ and ∞ elsewhere, $a = u(u-1)^{-1}$, and $u = 2.08$. Note that the solution for the logit procedure, if it exists, satisfies the bound restrictions $0 \leq w_i \leq 0.065$ for all i . The logit procedure was introduced as a case 6 in Deville and Särndal (1992). As the upper bound for the weight, 0.065 was used so that 3,026 samples (approximately 10%) have at least one raking regression weight greater than 0.065. In 99 samples among 30,000, no solution for the quadratic programming and logit procedure is possible because no feasible point satisfies (4) and the bound restriction. For those 99 samples, the maximum of the OLS regression weights was used as the upper bound for the quadratic programming and logit procedures.

Table 2 shows the average of the sum of squares for the six weights. The average weight is $1/32 = 0.03125$ for every estimator. The least squares procedures have the smallest sum of squares of the weights because this is the objective function for those procedures. The least squares procedures also have a slightly smaller range in the sum of squares. One percent of the least squares samples have a normalized mean of squares greater than 1.401 while one percent of the mean of squares for raking are greater than 1.441.

Table 2
Monte Carlo Average of the Sum of Squares of the Weights

	OLS	QP	SCW Reg	MLE Reg	Raking Reg	Logit
Average of $\mathbf{w}'\mathbf{w}$ ($\times 32$)	1.043	1.044	1.045	1.053	1.045	1.045

Table 3 contains properties for the minimum of the weights. Maximum likelihood has the largest average minimum weight while the least squares procedures have a smaller average for the minimum weight. The variance of the minimum weight is largest for the ordinary least squares procedures. Note that QP permits weights that equal the lower bound of zero.

Table 3
Monte Carlo Mean, Variance and Quantiles of the Minimum Weight

Procedure	Mean ($\times 10^2$)	Variance ($\times 10^5$)	Quantiles ($\times 32$)					
			0.01	0.10	0.50	0.90	0.99	
OLS	2.22	6.46	-0.10	0.34	0.79	0.96	1.00	
QP	2.21	6.32	0.00	0.32	0.79	0.96	1.00	
SCW Reg	2.44	3.58	0.22	0.49	0.84	0.97	0.99	
MLE	2.45	2.79	0.33	0.52	0.83	0.97	1.00	
Raking Reg	2.36	3.81	0.20	0.45	0.81	0.97	1.00	
Logit	2.25	5.23	0.09	0.36	0.78	0.96	1.00	

Among the procedures without bound restrictions on the weights, the ordinary least squares procedure has smaller maximum weight on average and much smaller variance for the maximum. See Table 4. The SCW-weighted regression has a smaller fraction of very large weights than MLE or raking ratio but a higher fraction of large weights than the ordinary least squares procedure. The bounded QP and

Logit procedures have smaller mean and variance for the maximum weight than the procedures with no upper bound restrictions.

Table 4
Monte Carlo Mean, Variance and Quantiles of the Maximum Weight

Procedure	Mean ($\times 10^2$)	Variance ($\times 10^5$)	Quantiles ($\times 32$)					
			0.01	0.10	0.50	0.90	0.99	
OLS	4.25	17.35	1.00	1.03	1.20	1.92	2.93	
QP	4.17	11.91	1.00	1.03	1.20	1.92	2.08	
SCW Reg	4.56	26.42	1.03	1.07	1.27	2.12	3.47	
MLE	4.75	56.13	1.00	1.04	1.25	2.31	4.72	
Raking Reg	4.46	30.25	1.00	1.03	1.23	2.09	3.63	
Logit	4.13	10.23	1.00	1.03	1.21	1.82	2.08	

To evaluate the performance of the procedures when the linear model does not hold, we considered estimation of the percentiles of the distribution function of x . Table 5 contains the Monte Carlo bias of the percentile estimators where the table entries are

$$[\min\{P, (1-P)\}]^{-1}[\hat{E}\{\hat{P}\} - P] \times 100,$$

and P is the percentile. For example, the Monte Carlo estimated relative bias in the ordinary least squares estimator of the 0.01 percentile is -7.75%. The ordinary least squares estimator has the largest biases in estimating the population percentiles, among the procedures without bound restrictions. The MLE has the smallest bias for all percentiles except the 75th, 95th and 99th, where the SCW-weighted regression estimator has the smallest bias. For samples of size 32, many samples contain no observation greater than the 99th percentile. The QP and Logit procedures have larger bias than other procedures except for the 75th percentile. The biases of the QP and Logit procedures are relatively large for the lower percentiles.

Table 5
Monte Carlo Standardized Bias in Percentile Estimators

Percentile	Procedure					
	OLS	QP	SCW Reg	MLE	Raking Reg	Logit
0.01	-7.75	-8.43	-2.88	-2.13	-4.70	-8.30
0.05	-7.27	-7.95	-2.58	-1.82	-4.30	-7.85
0.10	-6.66	-7.31	-2.27	-1.57	-3.91	-7.26
0.25	-5.25	-5.82	-1.79	-1.25	-3.13	-5.89
0.50	-3.21	-3.46	-1.37	-1.16	-2.18	-3.53
0.75	-2.30	-2.07	-1.60	-2.21	-2.25	-1.78
0.90	4.60	5.31	1.27	0.22	2.62	5.68
0.95	12.75	13.33	6.01	6.41	9.52	13.15
0.99	32.94	32.36	19.03	22.66	26.65	30.03

Table 6 contains the relative MSE of the percentile estimators where the table entries are

$$[\min\{P, (1-P)\}]^{-2}[\hat{E}\{\hat{P}\} - P]^2 \times 100.$$

Thus the relative mean square error of the OLS estimator of the 0.01 percentile is 283.27%. Although the OLS estimator

of the 0.01 percentile had the largest bias OLS has the smallest mean square error for the 0.01 percentile among the procedures without bound restrictions. The QP, OLS and Logit procedures are superior for the extreme percentiles while the other procedures perform better for the middle percentiles.

Table 6

Monte Carlo Relative MSE of Percentile Estimators							
Percentile	Procedure						
	OLS	QP	SCW Reg	MLE	Raking Reg	Logit	
0.01	283.27	282.50	309.23	311.58	296.37	282.76	
0.05	53.91	54.23	57.41	57.07	54.97	54.06	
0.10	25.50	25.97	26.40	25.79	25.26	25.80	
0.25	8.00	8.41	7.77	7.23	7.42	8.41	
0.50	1.99	2.07	1.88	1.71	1.83	2.12	
0.75	3.65	3.68	3.62	3.66	3.63	3.67	
0.90	14.50	14.60	14.25	14.57	14.36	14.56	
0.95	39.40	38.65	40.99	41.66	39.93	37.94	
0.99	200.17	196.24	235.71	216.22	205.85	194.33	

In 562 of 30,000 samples at least one of the OLS regression weights is negative. In 17 of the samples at least one of the original SCW regression weights was negative. The use of quadratic programming with the OLS objective function (QP) to produce weights greater than or equal to zero and less than 0.065 increases the average sum of squares by less than one percent. See Table 7. Using quadratic programming to bound the SCW regression weights (SCW (QPL)) by zero increases the average sum of squares very little because there are so few weights that are changed.

Table 7

Monte Carlo Average of the Sum of Squares of the Weights for Samples with at Least One Negative OLS Weight						
	SCW SCW				Raking	
	OLS	QP – Reg	(QPL)	MLE	– Reg	– Reg
Average of w'w (x32)	1.208	1.217	1.226	1.227	1.342	1.242

Table 8 gives the Monte Carlo MSE for the 562 samples with negative ordinary least squares weights. The quadratic programming procedure is superior to other nonnegative weight procedures for the 0.01 percentile and is inferior for the 0.99 percentile. Of the 562 samples, 497 had a sample mean greater than the population mean. Recall that the study population has an exponential distribution. Because the weight on the largest observation is zero in the 497 samples there is a 100 percent error in the quadratic programming estimator of the 0.99 percentile for most of the 497 samples with a sample mean greater than the population mean. In sampling from a finite population the bound on the weights would be greater than or equal to N^{-1} and the MSE of the quadratic programming procedure for the 0.99 percentile would be reduced.

Quadratic programming is superior to the other calibrated procedures for the 0.01 percentile in samples with negative

OLS weights. Raking regression and SCW-weighted regression are superior to MLE for the 0.01 and 0.05 percentiles. This is because MLE often has the largest maximum weight.

Table 8

Monte Carlo Relative MSE of Percentile Estimators for Samples with at Least One Negative OLS Weight						
Percentile	Procedure					
	OLS	QP	SCW (QPL)	MLE	Raking Reg	
0.01	287.52	291.11	350.58	461.80	344.06	
0.05	76.04	70.58	75.80	88.71	72.50	
0.10	44.80	40.74	39.31	38.84	36.05	
0.25	20.24	19.14	14.72	9.91	12.56	
0.50	5.03	5.31	3.65	2.26	3.35	
0.75	5.02	4.53	3.36	4.24	3.45	
0.90	23.77	23.69	20.04	18.80	20.49	
0.95	51.54	46.04	30.79	28.28	32.54	
0.99	206.33	90.08	39.40	57.54	43.49	

In 3,026 of 30,000 samples, at least one of the raking regression weights is greater than 0.065. In 2,152 samples, at least one of the OLS regression weights is greater than 0.065, and in 3,209 samples at least one of the SCW regression weights is greater than 0.065. The use of quadratic programming with the OLS objective function to produce weights in (0.000, 0.065) increases the average sum of squares by 1.5 percent. Using quadratic programming to bound the SCW regression weights by 0.000 and 0.065 increases the average sum of squares 0.8 percent. See the column for SCW (QP) of Table 9.

Table 9

Monte Carlo Average of the Sum of Squares of the Weights for Samples with at Least One Raking Reg Weight Greater than 0.065						
	SCW SCW Raking					
	OLS	QP – Reg	(QP) – Reg	Logit	MLE	
Average of w'w (x32)	1.210	1.228	1.221	1.231	1.228	1.290

Table 10 gives the Monte Carlo relative MSE for the 3,026 samples with raking regression weights greater than 0.065. The quadratic programming is superior to SCW (QP) and Logit for the 0.01, 0.95 and 0.99 percentile and the Logit procedure is superior to quadratic programming for other percentiles.

Table 10

Monte Carlo Relative MSE of Percentile Estimators for Samples with at Least One Raking Reg Weight Greater than 0.065							
Percentile	Procedure						
	OLS	QP	SCW Raking	– Reg	(QP) – Reg	Logit	MLE
0.01	139.96	130.53	173.86	146.40	124.02	173.65	206.65
0.05	39.83	42.88	39.35	41.69	39.87	37.14	40.83
0.10	26.31	30.92	22.40	28.10	28.88	20.21	19.98
0.25	13.56	17.72	10.13	15.69	17.71	8.65	7.01
0.50	3.95	4.87	3.32	4.75	5.37	3.03	2.28
0.75	4.84	5.35	4.89	5.58	5.37	5.05	5.48
0.90	27.98	29.04	28.70	29.34	29.32	28.79	32.07
0.95	74.15	67.54	85.02	68.12	65.98	83.13	95.99
0.99	198.77	179.58	219.16	181.17	172.45	212.38	226.73

Discussion

We began the research with the conjecture that starting with the SCW weights in a regression estimator would produce weights that were almost always positive and that the weights would have desirable properties as measured by the ability to estimate the distribution function of x . To some extent these results support the conjectures. The minimum weights of the SCW regression are larger than those of OLS and comparable to those for raking. Quadratic programming can be used to remove the negative weights in the few samples with negative weights. If no upper bound is imposed, the maximum weights for the SCW weighted regression fall between those of least squares and raking.

It is known that all of the procedures in our simulation study have the same order $n^{-1/2}$ properties. Our simulation and the study of generalized raking procedures done by Deville *et al.* (1993) indicate that there are also modest differences in small samples. No procedure is superior with respect to all criteria. Because of the poor performance for the extreme percentiles, we recommend against the use of the MLE objective function. The quadratic programming and Logit procedure produced weights with marginally smaller sums of squares, marginally smaller maximum weights, and marginally smaller MSE for extreme percentiles than the raking regression. The MLE, SCW regression and raking procedures give marginally larger minimum weights and marginally smaller MSE for the middle percentiles of the x distribution than quadratic programming and Logit procedure. The performances of quadratic programming and Logit procedures in estimating the distribution function of x are comparable.

Appendix

Proof. The ratio of the determinants of estimated covariance matrices in (20) is

$$\frac{\left| \tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)} \right|}{\left| \hat{\Sigma}_{\bar{x}\bar{x}} \right|} = 1 + O_p(n^{-1}) \quad (26)$$

by assumptions (24) and (25). The difference $\tilde{\mathbf{G}}_{\bar{x}\bar{x},(i)} - \hat{\mathbf{G}}_{\bar{x}\bar{x}}$ is

$$\begin{aligned} & (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N) \left(\tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} - \hat{\Sigma}_{\bar{x}\bar{x}}^{-1} \right) (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N)' \\ & - 2(\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N) \tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} \mathbf{d}'_{x_i} + \mathbf{d}_{x_i} \tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} \mathbf{d}'_{x_i}. \end{aligned}$$

By assumptions (23) and (24),

$$\begin{aligned} & \exp\{0.5[(\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N) (\tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} - \hat{\Sigma}_{\bar{x}\bar{x}}^{-1}) (\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N)']\} = \\ & 1 + O_p(n^{-1}). \quad (27) \end{aligned}$$

Using assumptions (24) and (19), the Taylor expansion at $\mathbf{d}_{x_i} = 0$ gives

$$\begin{aligned} & \exp[-(\bar{\mathbf{x}}_{\text{HT}} - \bar{\mathbf{x}}_N) \tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} \mathbf{d}'_{x_i} + 0.5 \mathbf{d}_{x_i} \tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} \mathbf{d}'_{x_i}] \\ & = 1 + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \tilde{\tilde{\Sigma}}_{\bar{x}\bar{x},(i)}^{-1} \mathbf{d}'_{x_i} + O_p(n^{-1}) \\ & = 1 + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\Sigma}_{\bar{x}\bar{x}}^{-1} \mathbf{d}'_{x_i} + O_p(n^{-1}). \quad (28) \end{aligned}$$

Thus, by (26), (27) and (28),

$$[N \tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-1}]^{-1} = (N \pi_i)^{-1} [1 + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\Sigma}_{\bar{x}\bar{x}}^{-1} \mathbf{d}'_{x_i}] + O_p(n^{-2}).$$

By assumptions (18), (23) and (25), and using the fact that $E\{\tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-2}\}$ is bounded,

$$\begin{aligned} \bar{y}_{p\tilde{\pi}} &= \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\boldsymbol{\theta}} + O_p(n^{-1}) \\ &= \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \boldsymbol{\theta}_N + O_p(n^{-1}). \quad (29) \end{aligned}$$

If one is an element of \mathbf{x}_i or $\text{Var}\{\sum_{i=1}^n \pi_i^{-1}\} = 0$, and if $\mathbf{M}_{\bar{x}\bar{y}} = \Sigma_{\bar{x}\bar{y}}$, the SCW-estimator for the population mean of vector $\mathbf{q}_i = (1, \mathbf{x}_i)$ satisfies

$$\bar{\mathbf{q}}_{p\tilde{\pi}} = N^{-1} \sum_{i=1}^n \tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-1} \mathbf{q}_i = (1, \bar{\mathbf{x}}_N) + O_p(n^{-1}), \quad (30)$$

because the $\boldsymbol{\theta}$ for \mathbf{x} is the identity matrix. By (30),

$$\begin{aligned} (\bar{\mathbf{x}}_c, \bar{y}_c) &= N \left[\sum_{i=1}^n \tilde{\pi}_{i|\bar{x}_{\text{HT}}}^{-1} \right]^{-1} (\bar{\mathbf{x}}_{p\tilde{\pi}}, \bar{y}_{p\tilde{\pi}}) \\ &= (\bar{\mathbf{x}}_{p\tilde{\pi}}, \bar{y}_{p\tilde{\pi}}) + O_p(n^{-1}). \quad (31) \end{aligned}$$

Thus,

$$\begin{aligned} \bar{y}_{w\text{reg}} &= \bar{y}_c + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_c) \hat{\boldsymbol{\beta}}_{c,1} \\ &= \bar{y}_{p\tilde{\pi}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{p\tilde{\pi}}) \hat{\boldsymbol{\beta}}_{c,1} + (\bar{y}_c - \bar{y}_{p\tilde{\pi}}) + (\bar{\mathbf{x}}_{p\tilde{\pi}} - \bar{\mathbf{x}}_c) \hat{\boldsymbol{\beta}}_{c,1} \\ &= \bar{y}_{p\tilde{\pi}} + O_p(n^{-1}) \\ &= \bar{y}_{\text{HT}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\text{HT}}) \hat{\boldsymbol{\theta}} + O_p(n^{-1}), \end{aligned}$$

by (30), (31) and (29).

Acknowledgements

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the USDA National Agricultural Statistics Service and the U.S. Bureau of the Census, and by Cooperative Agreement 68-3A75-14 between the USDA Natural Resources Conservation Service and Iowa State University. We thank the Associate editor and referees for comments that improved the paper.

References

- Bardsley, P., and Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 300-305.
- Husain, M. (1969). Construction of Regression Weights for Estimation in Sample Surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.
- Rao, J.N.K., and Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*, American Statistical Association, 57-64.
- Stephan, F.F. (1942). An alternative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.
- Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.
- Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities: complex design. *Survey Methodology*, 25, 57-66.

An Optimal Calibration Distance Leading to the Optimal Regression Estimator

Per Gösta Andersson and Daniel Thorburn¹

Abstract

When there is auxiliary information in survey sampling, the design based “optimal (regression) estimator” of a finite population total/mean is known to be (at least asymptotically) more efficient than the corresponding GREG estimator. We will illustrate this by some simulations with stratified sampling from skewed populations. The GREG estimator was originally constructed using an assisting linear superpopulation model. It may also be seen as a calibration estimator; *i.e.*, as a weighted linear estimator, where the weights obey the calibration equation and, with that restriction, are as close as possible to the original “Horvitz-Thompson weights” (according to a suitable distance). We show that the optimal estimator can also be seen as a calibration estimator in this respect, with a quadratic distance measure closely related to the one generating the GREG estimator. Simple examples will also be given, revealing that this new measure is not always easily obtained.

Key Words: Horvitz-Thompson estimator; Regression estimator; Survey sampling theory.

1. Notation and Basics

Consider a finite population U consisting of N objects labelled $1, \dots, N$ with associated study values y_1, \dots, y_N and J -dimensional auxiliary (column) vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$. We want to estimate the population total $t_y = \sum_{i \in U} y_i$ by drawing a random sample s of size n (fixed or random) from U , with first and second order inclusion probabilities $\pi_i = P(i \in s)$, $\pi_{ij} = P(i, j \in s)$, $i, j = 1, \dots, N$. The study values and the auxiliary vectors are recorded for the sampled objects and before the sample is drawn we assume that at least $t_x = \sum_{i \in U} \mathbf{x}_i$ is known.

This is the standard setup for a regression estimator. In section 2 we discuss different regression estimators: the common GREG estimator (Särndal, Swensson and Wretman 1992), the optimal estimator (Montanari 1987, Andersson, Nerman and Westhall 1995) and calibration estimators (Deville and Särndal 1992). It is well known that the GREG estimator can be obtained as a calibration estimator. In section 3 it is shown that this holds also for the optimal estimator, but with a more complicated distance measure. In the last two sections this and the optimal estimator are illustrated, first by theoretical examples and then by simulations.

Finally some comments about matrix notation in this paper: Generally, the transpose of a matrix A is denoted by A^T and if A is square, the inverse (generalised inverse) is written A^{-1} (A^-). We further let the column vectors $\mathbf{y} = (y_i)_{i \in s}$ and $\mathbf{w}_0 = (1/\pi_i)_{i \in s}$, \mathbf{X} be the $J \times n$ “design” matrix of the auxiliary information from s and finally \mathbf{I}_n means a unit diagonal matrix of size n .

2. Regression and Calibration Estimators

An unbiased simple estimator of t_y is the Horvitz-Thompson estimator $\hat{t}_y = \sum_{i \in s} y_i / \pi_i = \mathbf{y}^T \mathbf{w}_0$. However, more efficient estimators may be obtained utilising the auxiliary information, *e.g.*, the well-known model assisted GREG estimator, see Särndal *et al.* (1992). For example, constructed from the assumption of a homoscedastic linear regression superpopulation model the GREG estimator is

$$\hat{t}_{yr} = \mathbf{y}^T \mathbf{w}_0 + (\mathbf{y}^T \mathbf{R}_r \mathbf{X}^T) (\mathbf{X} \mathbf{R}_r \mathbf{X}^T)^{-1} (t_x - \hat{t}_x) \quad (1)$$

$$= \mathbf{y}^T \mathbf{g}, \quad (2)$$

where $\mathbf{R}_r = \mathbf{w}_0 \mathbf{I}_n$, $\hat{t}_x = \sum_{i \in s} \mathbf{x}_i / \pi_i$ and

$$\mathbf{g} = \left(\frac{1}{\pi_i} (1 + \mathbf{x}_i^T (\mathbf{X} \mathbf{R}_r \mathbf{X}^T)^{-1} (t_x - \hat{t}_x)) \right)_{i \in s}.$$

Now, the expression (2) for the GREG estimator is interesting since we also have that

$$\mathbf{x}^T \mathbf{g} = t_x, \quad (3)$$

which is called the *calibration equation*. This brings us to an alternative possible derivation of the GREG estimator according to Deville and Särndal (1992). Suppose that we seek an estimator $\mathbf{y}^T \mathbf{w}$ of t_y with a vector \mathbf{w} of sample-dependent weights $(w_i)_{i \in s}$, which respects the corresponding calibration equation, while also minimising the distance between \mathbf{w} and \mathbf{w}_0 according to the quadratic distance measure

1. Per Gösta Andersson, Mathematical Statistics, Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden; Daniel Thorburn, Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.

$$(\mathbf{w} - \mathbf{w}_0)^T \mathbf{R} (\mathbf{w} - \mathbf{w}_0),$$

where $\mathbf{R} = (\mathbf{w}_0 \mathbf{I}_n)^{-1}$.

This results in

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{R}^{-1} \mathbf{x}^T (\mathbf{X} \mathbf{R}^{-1} \mathbf{X}^T)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x), \quad (4)$$

which means that $\mathbf{w} = \mathbf{g}$, since here $\mathbf{R} = \mathbf{R}_r^{-1}$.

Turning to the optimal estimator, consider first the vector $(\hat{\mathbf{t}}_y, \hat{\mathbf{t}}_x^T)$ and let $\sum_{y,x}$ be the covariance (row) vector of $\hat{\mathbf{t}}_y$ and $\hat{\mathbf{t}}_x$ and $\sum_{x,x}$ the covariance matrix of $\hat{\mathbf{t}}_x$. Now, the minimum-variance, see Montanari (1987), unbiased linear estimator (in $\hat{\mathbf{t}}_y$ and $\hat{\mathbf{t}}_x$) of \mathbf{t}_y is the difference estimator

$$\hat{\mathbf{t}}_y + \sum_{y,x} \sum_{x,x}^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x). \quad (5)$$

Since $\sum_{y,x}$ and $\sum_{x,x}$ in practice are unknown, we let the optimal estimator be

$$\begin{aligned} \hat{\mathbf{t}}_{y\text{opt}} &= \mathbf{y}^T \mathbf{w}_0 + \hat{\sum}_{y,x} \hat{\sum}_{x,x}^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x) \\ &= \hat{\mathbf{t}}_y + (\mathbf{y}^T \mathbf{R}_{\text{opt}} \mathbf{X}^T) (\mathbf{X} \mathbf{R}_{\text{opt}} \mathbf{X}^T)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x), \end{aligned} \quad (6)$$

where $\mathbf{R}_{\text{opt}} = ((\pi_{ij} - \pi_i \pi_j) / (\pi_{ij} \pi_i \pi_j))_{i,j \in s}$.

In an asymptotic context, where $n \rightarrow \infty$ and $N \rightarrow \infty$, $\hat{\sum}_{x,y}$ and $\hat{\sum}_{x,x}$ may be viewed as components of the asymptotic covariance matrix of $(\hat{\mathbf{t}}_y, \hat{\mathbf{t}}_x^T)$. Under the assumption of consistency of $\hat{\sum}_{x,y}$ and $\hat{\sum}_{x,x}$, which holds under very mild conditions, see Andersson *et al.* (1995), the optimal estimator has the same asymptotic variance as the difference estimator (5). In particular it follows that the optimal estimator is asymptotically better than the usual GREG estimator, see Rao (1994), Montanari (2000) and Andersson (2001), *i.e.*, its asymptotic variance is never larger and usually smaller. In section 5 we actually present some simple simulations showing that the optimal estimator can be much more efficient than GREG. However, one does not know anything about the efficiency for finite samples, since the covariance estimator may converge slowly. The rate of convergence is illustrated in section 5. Note also that in some cases there exist asymptotically even better estimators which are not linear.

Now, the fact that the GREG estimator is also a calibration estimator using

$$(\mathbf{w} - \mathbf{w}_0)^T \mathbf{R}_r^{-1} (\mathbf{w} - \mathbf{w}_0) \quad (7)$$

as the distance measure and comparing (1) with (6), leads one to believe that replacing \mathbf{R}_r by \mathbf{R}_{opt} in (7) should imply that we instead derive the optimal regression estimator as a calibration estimator. That this actually holds is shown below.

3. The Main Result

In order to show existence of a distance measure corresponding to the optimal estimator, we will first state and prove a result in the general case.

Lemma: With \mathbf{R} denoting an arbitrary positive definite $n \times n$ matrix,

$$(\mathbf{w} - \mathbf{w}_0)^T \mathbf{R} (\mathbf{w} - \mathbf{w}_0) \quad (8)$$

subject to the constraint $\mathbf{X} \mathbf{w} = \mathbf{t}_x$, is minimised by

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{R}^{-1} \mathbf{X}^T (\mathbf{X} \mathbf{R}^{-1} \mathbf{X}^T)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x).$$

Proof: Introducing the $J \times 1$ vector $\boldsymbol{\lambda}$ of Lagrange multipliers, we get after differentiation the equation system

$$2\mathbf{R}(\mathbf{w} - \mathbf{w}_0) + \mathbf{X}^T \boldsymbol{\lambda} = 0 \quad (9)$$

$$\mathbf{X} \mathbf{w} - \mathbf{t}_x = 0 \quad (10)$$

Multiplying (9) by $\mathbf{X} \mathbf{R}^{-1}$, using (10) and solving for $\boldsymbol{\lambda}$, yields with $\mathbf{X} \mathbf{w}_0 = \hat{\mathbf{t}}_x$:

$$\boldsymbol{\lambda} = 2(\mathbf{X} \mathbf{R}^{-1} \mathbf{X}^T)^{-1} (\hat{\mathbf{t}}_x - \mathbf{t}_x). \quad (11)$$

Putting this into (9) and solving for \mathbf{w} finally leads to

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{R}^{-1} \mathbf{X}^T (\mathbf{X} \mathbf{R}^{-1} \mathbf{X}^T)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x).$$

From the lemma we thus have the following main result:

Theorem: With \mathbf{R}_{opt} being positive (semi-) definite and using the optimal calibration distance-measure, which we get by letting $\mathbf{R} = \mathbf{R}_{\text{opt}}^{-1} (\mathbf{R}_{\text{opt}}^{-1})$ in (8), the calibration estimator will become the optimal regression estimator.

Remark: \mathbf{R}_{opt} may in some cases be indefinite (see below). The only thing we know is that it is an unbiased estimator of a covariance matrix. If it is not positive semi-definite there also exist x -values such that $\mathbf{X} \mathbf{R}_{\text{opt}} \mathbf{X}^T$ is not positive semi-definite, but the probability of such x -values goes to zero as the population and sample sizes increase (and if $\sum_{x,x}$ is positive definite). A strict minimisation of a distance with "a negative component" would lead to infinitely large corrections. This problem of the optimal estimator has, to our knowledge, not been pointed out previously.

The simplest way to find a distance which gives the optimal estimator as a calibration estimator is to find a matrix \mathbf{R}_{dist} which has the same eigenvectors as \mathbf{R}_{opt} but where the eigenvalues are replaced by their absolute values. (This result can be shown along the same lines as the proof of the lemma above. The distance can be seen as the sum of

the products of the eigenvalues and the squared eigenvectors. Putting the derivatives equal to zero means that in the proposition we found the extremes *i.e.*, the minima for positive eigenvalues and the maxima for negative eigenvalues. By changing all negative signs the extremes will all be minima).

4. Examples

Positive definite \mathbf{R}_{opt} : Suppose that the objects in U are independently drawn with inclusion probabilities π_1, \dots, π_N (Poisson sampling); thus implying a random sample size n , where $E[n] = \sum_{i \in U} \pi_i$. Due to the independence of drawings, \mathbf{R}_{opt} is diagonal and specifically

$$\mathbf{R}_{\text{opt}}^{-1} = \mathbf{I}_n \begin{pmatrix} \pi_1^2 \\ \vdots \\ \pi_i^2 \\ \vdots \\ 1 - \pi_i \end{pmatrix}_{i \in s}.$$

Positive semi-definite \mathbf{R}_{opt} : Suppose n objects are drawn according to simple random sampling, *i.e.*, each object has inclusion probability $\pi_i = n/N$. The elements of \mathbf{R}_{opt} are

$$i = j: \left(\frac{N}{n} \right)^2 \frac{N - n}{N}$$

$$i \neq j: \left(\frac{N}{n} \right)^2 \frac{n - N}{N(n - 1)}.$$

This means that \mathbf{R}_{opt} is singular with rank $n - 1$.

Suppose instead (as in the following simulation study) that U is partitioned into L strata of sizes N_1, \dots, N_L , from which we draw independent simple random samples of sizes n_1, \dots, n_L . The elements of \mathbf{R}_{opt} then are

$$i = j: \left(\frac{N_h}{n_h} \right)^2 \frac{N_h - n_h}{N_h}$$

$$i \neq j: \left(\frac{N_h}{n_h} \right)^2 \frac{n_h - N_h}{N_h(n_h - 1)},$$

when in the latter case i and j both belong to stratum h , $h = 1, \dots, L$ and 0 otherwise. Therefore \mathbf{R}_{opt} has rank $N - h$.

Non positive semi-definite \mathbf{R}_{opt} : Let U consist of four elements and s of two elements. Suppose that a systematic sample is taken with probability 0.94 and a simple random sample with probability 0.06, *i.e.*, $\pi_{13} = \pi_{24} = 0.48$ and $\pi_{12} = \pi_{14} = \pi_{23} = \pi_{34} = 0.01$. In that case

$$\mathbf{R}_{\text{opt}} = \begin{pmatrix} 2 & 23/12 \\ 23/12 & 2 \end{pmatrix} \quad (12)$$

with probability 0.96 and

$$\mathbf{R}_{\text{opt}} = \begin{pmatrix} 2 & -96 \\ -96 & 2 \end{pmatrix} \quad (13)$$

with probability 0.04. The second matrix has a negative eigenvalue.

The problem does not necessarily disappear if N is large. Consider instead a population consisting of $N/4$ strata with four elements each. Suppose that the above sampling procedure is used independently in each stratum. In that case \mathbf{R}_{opt} will consist of a matrix with the above 2×2 – matrices along the diagonal and zeroes elsewhere.

5. A Simulation Study

5.1 Notation and Outline

In order to make empirical comparisons between the optimal estimator (OPT) and the GREG estimator (GREG) and also compare these estimators with the Horvitz-Thompson estimator (HT), we have conducted a small simulation study. In the previous sections we mentioned that OPT is Best Linear Asymptotic Efficient and a calibration estimator. Even though it has many nice properties it may for reasonable sample sizes be inefficient. Here we will in some simulated situations show that the optimal estimator can be a substantial improvement compared to GREG also for moderate sample sizes when the population is (deliberately) chosen to be unfavourable for GREG. A simple but non-trivial situation for which OPT is not equal to GREG arises for stratified simple random sampling, in particular, when the slopes differ between the different strata and the unstratified population. Consider therefore a population of size N , which is partitioned into L strata of sizes N_1, \dots, N_L . From each stratum h a simple random sample s_h of size n_h is drawn, where $s_1 + \dots + s_L = s$ and $n_1 + \dots + n_L = n$. For simplicity we further assume that the auxiliary information is one-dimensional and global, *i.e.*, only t_x is known beforehand. For GREG we have chosen the homoscedastic simple linear regression model, see Särndal *et al.* (1992).

The resulting expressions for HT, OPT and GREG respectively are

$$\hat{t}_y = N \bar{y}_{st}$$

$$\hat{t}_{y \text{ opt}} = N (\bar{y}_{st} + \hat{B}_{\text{opt}} (\bar{x} - \bar{x}_{st}))$$

$$\hat{t}_{yr} = N (\bar{y}_{st} + \hat{B}_r (\bar{x} - \bar{x}_{st})),$$

where $\bar{x} = (1/N) \sum_{i=1}^N x_i$, $\bar{y}_{st} = (1/N) \sum_{h=1}^L N_h \bar{y}_{s_h}$, (\bar{x}_{st} analogous) and

$$\hat{B}_{\text{opt}} = \frac{\sum_{h=1}^L \frac{N_h^2}{n_h-1} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sum_{i \in s_h} (x_i - \bar{x}_{s_h})(y_i - \bar{y}_{s_h})}{\sum_{h=1}^L \frac{N_h^2}{n_h-1} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sum_{i \in s_h} (x_i - \bar{x}_{s_h})^2}$$

$$\hat{B}_r = \frac{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i \in s_h} (x_i - \bar{x}_{st})(y_i - \bar{y}_{st})}{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i \in s_h} (x_i - \bar{x}_{st})^2}.$$

It is easily seen from these formulae that the optimal regression coefficient is the mean of the within stratum slopes and that the GREG regression coefficient is the global slope. When there is a large difference between these slopes the GREG correction becomes bad. We are here particularly interested in comparing the qualities of these estimators when the assisting (linear) model for GREG fails. We have thus generated x - and y -values from correlated lognormally distributed random variables X and Y , where $\ln X$ is normally distributed with expectation 0 and variance σ_1^2 ($N(0, \sigma_1^2)$) and $\ln Y$ is $N(0, \sigma_2^2)$. The variances σ_1^2 and σ_2^2 and the correlation between $\ln X$ and $\ln Y$ can then be chosen to obtain prespecified values of the variances σ_x^2 of X and σ_y^2 of Y and their correlation $\rho(X, Y)$. Values generated from bivariate normal distributions were obtained by MATLAB (version 6.0). Twelve populations have in this manner been created, each of size $N = 10,000$, including four combinations of variances σ_x^2 and σ_y^2 (10 and 100) and three values of the correlation $\rho(X, Y)$ (0.5, 0.7 and 0.9). For these populations a variance of 10 implies a skewness of 9.37 and the variance 100 leads to skewness 38.59.

Now, before stratification, the objects of each population are ordered with respect to ascending y -values. The number of strata is $L=5$ throughout with sizes $N_1 = 4,000$, $N_2 = 2,500$, $N_3 = 2,000$, $N_4 = 1,000$ and $N_5 = 500$. These strata are constructed in such a way that objects with the smallest y -values constitute stratum 1, and so forth. From each stratified population we have drawn samples of sizes $n = 250$, 1,000 and 2,500, where for each sample $n_1 = \dots = n_5$. This means that we have created an approximate π_{ps} (probability proportional to size) design, with for example, objects in stratum 5 having the largest inclusion probability (n_5/N_5). The number of simulated samples was $K = 25,000$ for each of the $12 \times 3 = 36$ cases and HT, OPT and GREG were then computed for each sample.

In general, common measures of quality for an estimator \hat{t} of a total t from a sequence $\hat{t}_1, \dots, \hat{t}_L$ are the estimated relative bias

$$\frac{\bar{\hat{t}} - t}{t}$$

and the estimated variance

$$S^2 = \frac{1}{K-1} \sum_{i=1}^K (\hat{t}_i - \bar{\hat{t}})^2,$$

where $\bar{\hat{t}} = (1/K) \sum_{i=1}^K \hat{t}_i$.

Since we are mainly concerned with comparisons of OPT and GREG, we will only display results of the relative measures of variance (or equivalently standard deviation)

$$\frac{S_{y \text{ opt}}^2}{S_{y \text{ HT}}^2} \quad \text{and} \quad \frac{S_{y r}^2}{S_{y \text{ HT}}^2},$$

from which we can compare the estimated variances of OPT and GREG with HT and also determine which of OPT and GREG have the lowest estimated variance.

5.2 Results

Firstly, as reference, the absolute value of the estimated relative bias of the unbiased HT did not in any case exceed $4 \cdot 10^{-4}$. The corresponding maximum values for OPT and GREG were $6 \cdot 10^{-3}$, which means that we may concentrate on the ratios of estimated variances in order to evaluate relative efficiencies of HT, OPT and GREG.

As seen from Table 1, OPT is superior to both HT and GREG (with one exception: $\rho(X, Y) = 0.9$, $\sigma_x^2 = 10$, $\sigma_y^2 = 100$ and $n = 250$, where GREG has slightly less estimated variance). For the lowest correlation though, the decrease in estimated variance for OPT compared with HT is not substantial. GREG on the other hand does not compete well with the others and this anomaly is particularly accentuated for the largest sample size $n = 2,500$. Changing $\rho(X, Y)$ to 0.7 means improvement for both OPT and GREG, but GREG is also now for most cases inferior to HT. Finally, for $\rho(X, Y) = 0.9$ GREG still displays poor behavior compared with HT for $n = 2,500$ (with the exception of $\sigma_x^2 = 100$ and $\sigma_y^2 = 10$). In general GREG is closing in on OPT for increasing values of $\rho(X, Y)$ (the assisting linear model becoming less misspecified), while OPT, on the other hand, is increasing its superiority over GREG for increasing sample sizes, which should come as no surprise since OPT is asymptotically well motivated.

Table 1
Relative Estimated Efficiencies (Given as Percentages) of OPT ($S^2_{y\text{ opt}}/S^2_{y\text{ HT}}$) and GREG ($S^2_{y\text{ r}}/S^2_{y\text{ HT}}$) to HT,
Based on 25,000 Simulated Samples for Each Sample Size

	$\sigma^2_x = 10$		$\sigma^2_x = 10$		$\sigma^2_x = 100$		$\sigma^2_x = 100$	
	$\sigma^2_y = 10$		$\sigma^2_y = 100$		$\sigma^2_y = 10$		$\sigma^2_y = 100$	
	OPT	GREG	OPT	GREG	OPT	GREG	OPT	GREG
$\rho(X, Y) = 0.5$								
$n = 250$	99.1	232.8	97.4	176.8	93.9	179.4	91.4	122.3
$n = 1,000$	98.3	247.1	98.0	193.7	97.5	183.5	99.9	141.9
$n = 2,500$	96.8	756.7	96.8	1,455.0	97.8	534.7	96.8	1,625.5
$\rho(X, Y) = 0.7$								
$n = 250$	89.7	197.6	83.8	101.2	73.6	120.4	64.3	72.9
$n = 1,000$	91.0	227.5	89.8	117.2	81.2	120.5	71.7	84.0
$n = 2,500$	93.8	648.2	91.5	1,308.6	93.1	218.6	93.1	673.5
$\rho(X, Y) = 0.9$								
$n = 250$	56.5	76.1	41.2	38.8	27.2	43.4	40.4	41.4
$n = 1,000$	61.8	87.3	44.1	44.2	27.6	44.1	41.5	45.4
$n = 2,500$	77.0	237.4	59.8	335.4	63.6	66.0	74.6	259.8

References

Andersson, P.G. (2001). Improving estimation quality in large sample surveys. Ph. D. Thesis, Department of Mathematics, Chalmers University of Technology and Göteborg University.

Andersson, P.G., Nerman, O. and Westhall J. (1995). Auxiliary information in survey sampling. *Technical Report NO 1995:3*, Department of Mathematics, Chalmers University of Technology and Göteborg University.

Deville, J.C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.

Montanari, G.E. (2000). Conditioning on auxiliary variable means in finite population inference. *Australian & New Zealand Journal of Statistics*, 42, 407-421.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

Approximations to b^* in the Prediction of Design Effects Due to Clustering

Peter Lynn and Siegfried Gabler¹

Abstract

Kish's well-known expression for the design effect due to clustering is often used to inform sample design, using an approximation such as \bar{b} in place of b . If the design involves either weighting or variation in cluster sample sizes, this can be a poor approximation. In this article we discuss the sensitivity of the approximation to departures from the implicit assumptions and propose an alternative approximation.

Key Words: Complex sample design; Intraclass correlation coefficient; Selection probabilities; Weighting.

1. Alternative Functions of Cluster Size

Kish (1965) used an expression for the design effect (variance inflation factor) due to sample clustering, $\text{deff} = 1 + (b - 1) \rho$, where b is the number of observations in each cluster (primary sampling unit) and ρ is the intraclass correlation coefficient. This expression is well-known, is taught on courses on sampling theory, and is used by survey practitioners in designing and evaluating samples.

The expression holds when there is no variation in cluster sample size and the design is equal-probability (self-weighting). We can express these two criteria formally:

$$b_c = b \quad \forall c \quad (1)$$

where $c = 1, \dots, C$ denote the clusters, and

$$w_i = w \quad \forall i \quad (2)$$

where $i = 1, \dots, I$ denote the weighting classes, with w_i the associated design weights.

However, most surveys involve departures from (1) and (2). In the general case, *i.e.*, removing restrictions (1) and (2), Gabler, Häder and Lahiri (1999) showed that under an appropriate model, $\text{deff}_c = 1 + (b^* - 1) \rho$, where

$$b^* = \frac{\sum_{c=1}^C \left(\sum_{i=1}^I w_i b_{ci} \right)^2}{\sum_{i=1}^I w_i^2 b_i} = \frac{\sum_{c=1}^C \left(\sum_{j=1}^{b_c} w_{cj} \right)^2}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}^2} \quad (3)$$

and b_{ci} is the number of observations in weighting class i in cluster c , $b_i = \sum_{c=1}^C b_{ci}$ (we have changed the notation from that of Gabler *et al.* (1999), to provide consistency) and w_{cj} is the weight associated with the j^{th} observation in cluster c , $j = 1, \dots, b_c$.

The quantity b^* can be calculated from survey micro-data, provided the design weight and cluster membership is known for each observation. However, at the sample design stage it is not clear how b^* can be predicted. Gabler *et al.*

(1999) interpreted Kish's b as a form of weighted average cluster size:

$$\begin{aligned} \bar{b}_w &= \frac{\sum_{c=1}^C b_c \left(\sum_{i=1}^I w_i^2 b_{ci} \right)}{\sum_{c=1}^C \sum_{i=1}^I w_i^2 b_{ci}} \\ &= \frac{\sum_{c=1}^C \left(b_c \sum_{j=1}^{b_c} w_{cj}^2 \right)}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}^2} \end{aligned} \quad (4)$$

where b_c is the number of observations in cluster c , $b_c = \sum_{i=1}^I b_{ci}$. However, (4) is no easier than (3) to predict at the sample design stage. A simpler interpretation, perhaps commonly used in sample design, is the unweighted mean cluster size:

$$\bar{b} = \frac{\sum_{c=1}^C b_c}{C} = m/C. \quad (5)$$

It is much easier to predict \bar{b} at the sample design stage than either \bar{b}_w or b^* , as it requires knowledge only of the total number of observations, m , and total number of clusters, C .

2. Relationship Between b^* , \bar{b}_w and \bar{b} Under Alternative Assumptions

Let

$$\bar{w}_c = \frac{1}{b_c} \sum_{j=1}^{b_c} w_{cj} = \sum_{i=1}^I w_i \frac{b_{ci}}{b_c},$$

$$\text{Cov}(b_c, b_c \bar{w}_c^2) = \frac{1}{C} \sum_{c=1}^C b_c^2 \bar{w}_c^2 - \frac{m}{C^2} \sum_{c=1}^C b_c \bar{w}_c^2$$

and

1. Peter Lynn, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom. E-mail: p.lynn@essex.ac.uk; Siegfried Gabler, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, 68072 Mannheim, Germany. E-mail: gabler@zuma-mannheim.de.

$$\begin{aligned}\text{Var}(w_{cj}) &= \frac{1}{b_c} \sum_{j=1}^{b_c} (w_{cj} - \bar{w}_c)^2 \\ &= \sum_{i=1}^I \frac{b_{ci}}{b_c} (w_i - \bar{w}_c)^2 \quad \forall c.\end{aligned}$$

Then

$$b^* = \frac{C \cdot \text{Cov}(b_c, b_c \bar{w}_c^2) + \bar{b} \sum_{c=1}^C b_c \bar{w}_c^2}{\sum_{c=1}^C b_c \cdot \text{Var}(w_{cj}) + \sum_{c=1}^C b_c \bar{w}_c^2}. \quad (6)$$

If (1) holds, then (6) becomes:

$$b^* = \bar{b} \left(\frac{\sum_{c=1}^C \bar{w}_c^2}{\sum_{c=1}^C \text{Var}(w_{cj}) + \sum_{c=1}^C \bar{w}_c^2} \right). \quad (7)$$

So, in that circumstance, $b^* \leq \bar{b}$. If, additionally, weights are equal within clusters, viz:

$$w_{cj} = w_c \quad \forall j \in c \quad (8)$$

then $b^* = \bar{b}$.

If (8) holds, but not (1), then

$$b^* \geq \bar{b} \quad \text{if and only if} \quad \text{Cov}(b_c, b_c \bar{w}_c^2) \geq 0$$

$$\text{since } b^* - \bar{b} = \frac{C \cdot \text{Cov}(b_c, b_c \bar{w}_c^2)}{\sum_{c=1}^C b_c \bar{w}_c^2}.$$

The covariance would be negative only if small cluster sizes coincide with large average weights within the clusters and *vice versa*. In section 4 below, we observe that this did not occur in any country on round 1 of the European Social Survey. Furthermore, from (3) and (4), we have:

$$b^* = \bar{b}_w = \sum_{c=1}^C (w_c b_c)^2 / \sum_{c=1}^C w_c^2 b_c. \quad (9)$$

If we additionally impose the restriction (1), then we have the obvious result $b^* = \bar{b}_w = \bar{b} = b_c \bar{w}_c$.

The result in (9) would apply to surveys where the only variation in selection probabilities was due to disproportionate sampling between domains that did not cross-cut clusters. A common example would involve disproportionate stratification by region, with PSUs consisting of geographical areas hierarchical to regions.

A practical relaxation of the restriction on the variation in weights is:

$$b_{ci} = b_c \left(\frac{b_i}{m} \right) \quad \forall i, c. \quad (10)$$

In other words, we allow variation in weights within clusters, but we constrain the weights to have the same relative frequency distribution in each cluster, *i.e.*, the means and the variances of the weights within clusters do not depend on the clusters.

Now, (3) simplifies as follows:

$$\begin{aligned}b^* &= \sum_{c=1}^C \left(\sum_{i=1}^I w_i b_c \frac{b_i}{m} \right)^2 / \sum_{i=1}^I w_i^2 b_i \\ &= \sum_{c=1}^C \left(b_c^2 \left(\sum_{i=1}^I w_i b_i \right)^2 \right) / m^2 \sum_{i=1}^I w_i^2 b_i \\ &= \frac{\left(\sum_{i=1}^I w_i b_i \right)^2 \sum_{c=1}^C b_c^2}{\sum_{i=1}^I w_i^2 b_i m^2}.\end{aligned} \quad (11)$$

Note that $((\sum_{i=1}^I w_i b_i)^2) / \sum_{i=1}^I w_i^2 b_i = m / (1 + c_w^2)$, where c_w^2 is the squared coefficient of variation, across all observations, of the weights. Also, $(\sum_{c=1}^C b_c^2) / m^2 = (1 + c_b^2) / C$, where c_b^2 is the squared coefficient of variation, across all clusters, of the cluster sample sizes. Thus, (11) becomes:

$$b^* = \frac{m}{(1 + c_w^2)} \frac{(1 + c_b^2)}{C} = \bar{b} \frac{(1 + c_b^2)}{(1 + c_w^2)} = \bar{b}, \text{ say.} \quad (12)$$

So, \bar{b} will underestimate b^* if $c_b^2 > c_w^2$ and *vice versa*. In particular, if $w_{cj} = w \forall j, c$ and $c_b^2 > 0$, then $b^* > \bar{b}$. The greater the variation in b_c , the greater the extent to which \bar{b} will under-estimate b^* .

Assumption (10) will rarely hold exactly, but this result might be useful in situations where the distribution of weights is expected to be similar across clusters. An example might be address-based samples where one person is selected per address. If the distribution of the number of persons per address is approximately constant across PSUs (in the population), then the distribution of weights will vary across clusters in the sample only due to sampling variation and disproportionate nonresponse (the effect of this could, of course, be substantial if cluster sample sizes are small).

If no restriction is imposed on the variation in weights, but $\text{Var}(w_{cj}) > 0$ for at least one c , then, from (6),

$$b^* \geq \bar{b} \quad \text{if and only if} \quad \zeta = \frac{C^2 \text{Cov}(b_c, b_c \bar{w}_c^2)}{m \sum_{c=1}^C b_c \text{Var}(w_{cj})} \geq 1. \quad (13)$$

If (10) holds, then $\zeta = c_b^2 / c_w^2$.

3. Implications for Sample Design

Expression (12) suggests that b^* may be predicted by predicting the relative magnitudes of c_b^2 and c_w^2 . However, this result applies to a special situation, where

$$\begin{aligned}\text{Cov}(w_{cj}, b_c) &= \frac{1}{m} \sum_{c=1}^C \sum_{j=1}^{b_c} (w_{cj} - \bar{w}) (b_c - \bar{b}) \\ &= \frac{1}{m} \sum_{c=1}^C (b_c - \bar{b}) \left(\sum_{i=1}^I w_i b_{ci} - b_c \bar{w} \right) \\ &\stackrel{\text{from (10)}}{=} \frac{1}{m^2} \sum_{c=1}^C (b_c - \bar{b}) b_c \left(\sum_{i=1}^I w_i b_i - m \bar{w} \right) \\ &= 0\end{aligned}$$

where

$$\begin{aligned}\bar{w} &= \frac{1}{m} \sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj} = \frac{1}{m} \sum_{c=1}^C b_c \bar{w}_c \\ \bar{b} &= \frac{1}{m} \sum_{c=1}^C \sum_{j=1}^{b_c} b_c = \frac{1}{m} \sum_{c=1}^C b_c^2 = \frac{m}{C} (1 + c_b^2).\end{aligned}$$

When this covariance is expected to be small, it may be appropriate to predict b^* thus:

$$b^* = \hat{b} = \hat{b} \frac{(1 + \hat{c}_b^2)}{(1 + \hat{c}_w^2)}.\tag{14}$$

Both coefficients of variation can be estimated from knowledge of the proposed sample design. In the following section, we investigate sensitivity of predictions obtained in this way to assumption (10) using real data from different sample designs with $\text{Cov}(w_{cj}, b_c) > 0$.

4. Example: European Social Survey

The European Social Survey (ESS) is a cross-national survey for which great efforts have been made to achieve approximate functional equivalence in sample design between participating nations (Lynn, Häder, Gabler and Laaksonen 2004). Nevertheless, there is considerable variety in the types of design used, primarily due to variation in the nature of available frames and in local objectives, such as a desire for sub-national analysis which may lead to disproportionate stratification by domain. We use here data from the first round of the ESS, for which fieldwork was carried out in 2002 – 2003. Of the 22 participating nations, 17 had a clustered sample design. Of these, two had not yet provided useable sample data at the time of writing. In Table 1 we present the sample values of b^* , \bar{b} , c_b^2 , c_w^2 , \bar{b} , $|\bar{b} - b^*|$, $|\bar{b} - b^*|$, $\text{Corr}(w_{cj}, b_c)$, and ζ for the remaining 15. Note that the United Kingdom and Poland both had a 2 – domain design with the sample clustered only in one domain, namely Great Britain (*i.e.*, excluding Northern Ireland) and less densely-populated areas (*i.e.*, all except the largest 42 towns) respectively. Figures presented in table 1 relate only to the clustered domain.

Table 1
Sample Values of b^* , \bar{b} , c_b^2 , c_w^2 , \bar{b} , $|\bar{b} - b^*|$, $|\bar{b} - b^*|$, $\text{Corr}(w_{cj}, b_c)$, and ζ , for 15 Surveys

Country		b^*	\bar{b}	c_b^2	c_w^2	\bar{b}	$ \bar{b} - b^* $	$ \bar{b} - b^* $	$\text{Corr}(w_{cj}, b_c)$	ζ
Austria	AT	6.49	7.08	0.08	0.25	6.15	0.34	0.58	0.0036	0.4549
Belgium	BE	6.56	5.79	0.13	0.00	6.56	0.00	0.77	.	.
Switzerland	CH	8.83	9.23	0.12	0.21	8.50	0.34	0.40	0.0223	0.7060
Czech Republic	CZ	2.94	2.70	0.24	0.25	2.68	0.26	0.24	0.0225	1.7350
Germany	DE	18.85	18.13	0.07	0.11	17.42	1.43	0.72	-0.2287	.
Spain	ES	4.96	5.04	0.17	0.22	4.80	0.15	0.08	-0.0767	0.8757
Great Britain	GB	11.11	12.27	0.08	0.22	10.90	0.21	1.16	0.0114	0.4198
Greece	GR	5.47	5.86	0.09	0.22	5.25	0.22	0.39	-0.0280	0.5207
Hungary	HU	8.68	8.18	0.06	0.00	8.68	0.00	0.50	.	.
Ireland	IE	12.09	11.18	0.13	0.04	12.05	0.05	0.91	0.0006	3.1054
Israel	IL	11.79	12.82	0.12	0.56	9.27	2.53	1.02	-0.1271	0.4401
Italy	IT	10.98	10.87	0.26	0.16	11.80	0.83	0.10	-0.5589	1.3018
Norway	NO	44.09	18.68	1.33	0.01	43.32	0.77	25.41	0.0807	.
Poland (rural)	PL	10.07	9.45	0.06	0.01	9.88	0.19	0.62	0.2923	.
Slovenia	SI	10.76	10.13	0.06	0.00	10.76	0.00	0.63	.	.

From (12), we would expect to observe $\bar{b} > b^*$ when $\hat{c}_w^2 > \hat{c}_b^2$. A common sample design for which this inequality can be anticipated is one where, a) the selected cluster sample size is constant, so variation in b_c will be limited to that caused by differential non-response; and b) the samples are equal-probability samples of addresses, with subsequent random selection of one person per address, leading to variation in design weights reflecting the variation in household size. There are six nations with sample designs of this type (AT, CH, ES, GB, GR, IL). It is indeed the case that for all of these nations, $\zeta < 1$ and $\bar{b} > b^*$. Furthermore, for 5 of these 6 nations (AT, CH, ES, GB, GR, $h=1, \dots, 5$) we might expect (10) to be a reasonable approximation as the only variation in weights is that due to selection within a household/address. For these, we might expect \hat{b} to perform better than \bar{b} . Indeed, $|\bar{b} - b^*| < |\hat{b} - b^*|$ for 4 of the 5, and $(\sum_{h=1}^5 |\bar{b} - b^*|) / \sum_{h=1}^5 |\hat{b} - b^*| = 0.48$. The one nation where \bar{b} would not provide an improvement is Spain and this is to be expected as \bar{b} is small. Small cluster sample sizes leave them relatively more susceptible to the effects of nonresponse and also sampling variance, which will lead to violation of (10). In Israel, there was a further source of variation in design weights as there was disproportionate stratification by geographical areas. This too causes violation of (10), so we would not expect \hat{b} necessarily to provide an improvement on \bar{b} as a predictor of b^* .

Of the nations where $c_b^2 < c_w^2$, there is only one (CZ) for which $\bar{b} < b^*$ and $\zeta > 1$. This is also the nation with the smallest value of \bar{b} . When cluster sample sizes are particularly small, deff will be small and the choice between estimators of b^* may be less important.

There are five nations where sample units were individuals selected with equal probabilities (within clusters) from population registers (BE, DE, HU, PL, SI). In this case (8) (and, therefore, (10)) holds strictly, so we have $\bar{b} < b^*$. For three of these nations (BE, HU, SI) the sample is equal-probability, so we observe $\bar{b} = b^*$. It is clear that \hat{b} is superior to \bar{b} for equal-probability samples. For Germany and Poland, there is some variation in design weights between clusters (but not within). This variation is modest in Poland, and $|\bar{b} - b^*| < |\hat{b} - b^*|$, but the same is not true in Germany, where the ex-East Germany was sampled at a considerably higher rate than the ex-West Germany.

The Norwegian sample design was the only one that resulted in considerable variation in cluster sample sizes at the selection stage. The dramatic impact of this on $\bar{b} - b^*$ can clearly be seen. Again, this is a situation in which \hat{b} is likely to be preferable to \bar{b} as a predictor of b^* .

The designs in Ireland and Italy both involved selecting addresses from the electoral registers with probability

proportional to number of electors and then selecting one resident at random from each selected address. Such designs are not equal-probability, but are likely to result in considerably less variation in design weights than the address-based sample designs discussed earlier (Lynn and Pisati 2005). In both these cases, $\hat{c}_w^2 < \hat{c}_b^2$, the difference being greater in the case of Italy where some cluster sample sizes (in the largest municipalities) were considerably larger than the others (in Ireland, all were equal at the selection stage). Aside from the Czech Republic, these are the only two nations with $\zeta > 1$.

5. Conclusion

To aid prediction of the design effect due to clustering, we believe that \hat{b} is likely to be a better choice than \bar{b} as a predictor of b^* in situations where it can reasonably be expected that (10) will approximately hold. This includes, but is not restricted to, the following common types of sample design:

- Equal-probability designs where cluster sample sizes vary by design;
- Equal-probability designs where clusters do not vary by design but are likely to vary due to nonresponse;
- Address-based samples where one person is selected at each address, there is no other significant source of variation in selection probabilities, and cluster sizes do not vary by design.

Acknowledgement

This research was carried out while the first author was Guest Professor at the Center for Survey Research and Methodology (ZUMA), Mannheim, Germany.

References

- Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Lynn, P., Häder, S., Gabler, S. and Laaksonen, S. (2004). Methods for Achieving Equivalence of Samples in Cross-National Surveys. ISER Working Paper 2004-09. Available at <http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-09.pdf>.
- Lynn, P., and Pisati, M. (2005). Improving the quality of sample design for social surveys in Italy: Lessons from the European Social Survey. Forthcoming.

A Note on the C_p Statistic Under the Nested Error Regression Model

Jane L. Meza and P. Lahiri¹

Abstract

Nested error regression models are frequently used in small-area estimation and related problems. Standard regression model selection criterion, when applied to nested error regression models, may result in inefficient model selection methods. We illustrate this point by examining the performance of the C_p statistic through a Monte Carlo simulation study. The inefficiency of the C_p statistic may, however, be rectified by a suitable transformation of the data.

Key Words: C_p statistics; Nester error regression model; Monte Carlo simulation.

1. Introduction

This paper examines the limitations of a standard regression model selection criterion, C_p the statistic, for nested error regression models. The C_p statistic (Mallows 1973) is defined by

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} - n + 2p \quad (1)$$

where RSS_p is the residual sum of squares and p is the number of parameters for model P , n is the number of observations and $\hat{\sigma}^2$ is an estimate of σ^2 . If the model is correct, the value of C_p should be similar to or smaller than p . The C_p model selection criterion is sensitive to outliers and departures from the normal i.i.d. assumption on the errors. The C_p statistic therefore cannot be directly applied to the nested error regression model since here the error structure is not i.i.d.

We propose a transformation that adjusts for intraclass correlation and allows use of the standard C_p model selection criterion. The method presented in this paper can be applied to select covariates in the analysis of complex survey data and small-area models. For example, our technique could be used to select covariates in the nested error regression model used by Battese, Harter and Fuller (1988) to estimate the area planted (in hectares) with corn or soybeans for twelve Iowa counties. They used the following model:

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij}, \quad (2)$$

for unit $j = 1, \dots, n_i$ in county $i = 1, \dots, m$, where n_i is the sample size for small area i and the total sample size is $n = \sum_{i=1}^m n_i$. The county effects, v_i , are distributed as $N(0, \sigma_v^2)$ independent of the random errors e_{ij} , which are distributed as $N(0, \sigma_e^2)$. The area (in hectares) in unit j of county i is denoted by y_{ij} and $x_{ij} = (1, x_{ij1}, \dots, x_{ijp})$ is a

$p+1$ vector of the values of the covariates x_{i1}, \dots, x_{ip} for unit j in county i . The vector $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ is a $p+1$ vector of unknown parameters.

The nested error regression model can be expressed in matrix form as

$$y = X\beta + \varepsilon \quad (3)$$

where $y = (y'_1, \dots, y'_m)'$, $y'_i = (y_{i1}, \dots, y_{in_i})'$, $\varepsilon = (\varepsilon'_1, \dots, \varepsilon'_m)'$, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})'$, $\varepsilon_{ij} = v_i + e_{ij}$. Further, $X' = (X'_1, \dots, X'_m)$ where X_i is an $n_i \times (p+1)$ matrix with rows x_{ij} for $j = 1, \dots, n_i$, $\varepsilon \sim N(0, \sigma^2 V)$ where $\sigma^2 = \sigma_v^2 + \sigma_e^2$, V has block-diagonal form $\bigoplus_1^m V_i$ with $V_i = (1-\rho)I_{n_i} + \rho J_{n_i}$ where $\rho = \sigma_v^2 / \sigma^2$ is the common intrastratum correlation, I_{n_i} is the $n_i \times n_i$ identity matrix and J_{n_i} is the $n_i \times n_i$ unit matrix.

Since the nested error model does not have i.i.d errors, standard regression procedures do not apply. The simulation study in section 3 reveals that the C_p criterion does not perform well under the nested error regression model. The transformations considered in the next section are used to transform the nested error regression model into a standard regression model with i.i.d. errors. With these transformed observations, the C_p criterion performs much better.

2. Adjusting for Intra-area Correlations

As noted in the previous section, conventional model selection methods like the C_p criterion are not appropriate since the intrastratum correlations are ignored. Wu, Holt and Holmes (1988) and Rao, Sutradhar and Yue (1993) studied the effect of conventional methods for the nested error regression model in a different context.

Consider the nested error regression model and let $\sigma^2 = \sigma_v^2 + \sigma_e^2$ and ρ be the common intra-area correlation, $\rho = \sigma_v^2 / \sigma^2$. As in Fuller and Battese (1973) and Rao *et al.*

(1993), transform the nested error regression model into a standard regression model with i.i.d. errors.

Let

$$\alpha_i = 1 - \left[\frac{1 - \rho}{1 + (n_i - 1)\rho} \right]^{1/2}, \quad (4)$$

$$y_{ij}^* = y_{ij} - \alpha_i \bar{y}_i, \quad (5)$$

$$x_{ij}^* = x_{ij} - \alpha_i \bar{x}_i, \quad (6)$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ and $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i$. The transformed model then becomes

$$y_{ij}^* = x_{ij}^* \beta + e_{ij}^*, \quad (7)$$

for $j = 1, \dots, n_i, i = 1, \dots, m$ and e_{ij}^* are independently distributed as $N(0, \sigma_e^2)$. Now, the standard C_p model selection criterion may be applied to the transformed data.

In practice, ρ is usually unknown and must be estimated from the data. Rao *et al.* (1993) used Henderson's (1953) method to obtain unbiased quadratic estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ of the variance components σ_v^2 and σ_e^2 . Once the estimators have been obtained, $\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$ may be estimated by

$$\hat{\rho} = \max \left[0, \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2} \right]. \quad (8)$$

To obtain the estimators of the variance components, let $\{u_{ij}\}$ be the residuals from the ordinary least squares regression of $\{y_{ij} - \bar{y}_i\}$ on $\{x_{ij1} - \bar{x}_{i,1}, \dots, x_{ijp} - \bar{x}_{i,p}\}$ without the intercept term, where $x_{i,l} = \sum_{j=1}^{n_i} x_{ijl} / n_i$ for $l = 1, \dots, p$. Let $\{r_{ij}\}$ be the residuals from the ordinary

least squares regression of y_{ij} on $\{x_{ij0}, \dots, x_{ijp}\}$ with the intercept term.

The estimators of σ_v^2 and σ_e^2 are given by

$$\hat{\sigma}_e^2 = (n - m - p - 1 - \lambda)^{-1} \sum_{i=1}^m \sum_{j=1}^{n_i} e_{ij}^2, \quad (9)$$

$$\hat{\sigma}_v^2 = n_*^{-1} \left[\sum_{i=1}^m \sum_{j=1}^{n_i} r_{ij}^2 - (n - p - 1) \hat{\sigma}_e^2 \right], \quad (10)$$

$$n_* = n - \text{tr} \left[(X'X)^{-1} \sum_{i=1}^m n_i^2 \bar{x}_i \bar{x}_i' \right] \quad (11)$$

where $\lambda = 0$ if the model has no intercept term and $\lambda = 1$ otherwise. We propose to apply standard C_p model selection criterion on these transformed observations y_{ij}^* and x_{ij}^* .

3. A Simulation Study

A simulation study was conducted to examine the behavior of the C_p model selection criterion and the proposed transformations for the nested error regression model. The following model was considered:

$$y_{ij} = \beta_0 x_{ij0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + v_i + e_{ij} \quad (12)$$

for $i = 1, \dots, 10, n_i \in \{2, \dots, 5\}$, $j = 1, \dots, n_i$ and $n = 40$. The v_i are distributed as $N(0, \sigma_v^2)$ independent of e_{ij} which are distributed as $N(0, 1)$. The data x_{ijl} are taken from an example given by Gunst and Mason (1980) and included in Shao (1993) (Table 1). The value of x_{ij0} is 1 for all $i = 1, \dots, 10, j = 1, \dots, n_i$.

Table 1
Data for Nested Error Simulation

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0.3600	0.5300	1.0600	0.5326	0.0900	0.1800	0.5900	0.1855
1.3200	2.5200	5.7400	3.6183	0.0200	0.1600	0.2400	0.1572
0.0600	0.0900	0.2700	0.2594	0.0200	0.1100	0.2100	0.0998
0.1600	0.4100	0.8300	1.0346	0.0500	0.2400	0.4300	0.2804
0.0100	0.0200	0.0700	0.0381	0.1100	0.3900	0.2900	0.2879
0.0200	0.0700	0.0700	0.3440	0.1800	0.1100	0.4300	0.6810
0.5600	0.6200	2.1200	1.4559	0.0400	0.0900	0.2300	0.3242
0.9800	1.0600	2.8900	4.0182	0.8500	1.3300	2.7000	2.6013
0.3200	0.2000	0.7600	0.4600	0.1700	0.3200	0.6600	0.4469
0.0100	0.0000	0.0700	0.1540	0.0800	0.1200	0.4900	0.2436
0.1500	0.2500	0.5000	0.6516	0.3800	0.1800	0.4900	0.4400
0.2400	0.2800	0.5900	0.0611	0.1100	0.1300	0.1800	0.3351
0.1100	0.3500	0.4000	0.1922	0.3900	0.3800	0.9900	1.3979
0.0800	0.1300	0.2800	0.0931	0.4300	0.4600	1.4700	2.0138
0.6100	0.8500	0.4900	0.0538	0.5700	1.1600	1.8200	1.9356
0.0300	0.0300	0.2300	0.0199	0.1300	0.0300	0.0800	0.1050
0.0600	0.1100	0.5000	0.0419	0.0400	0.0500	0.1400	0.2207
0.0200	0.0800	0.2500	0.1093	0.1300	0.1800	0.2800	0.0180
0.0400	0.2400	0.0800	0.0328	0.2000	0.9500	0.4100	0.1017
0.0000	0.0200	0.0400	0.0797	0.0700	0.0600	0.1800	0.0962

Some of the β_k may be zero and thus various combinations of variables were chosen from $(x_0, x_1, x_2, x_3, x_4)$ to be the predictors used to generate data coming from a nested error regression model. There are $2^p - 1 = 31$ possible models. Each model will be denoted by a subset of $(0, 1, 2, 3, 4)$ that contains the indices of the variables x_i in the model.

Data were generated using 1,000 simulations for several values of σ_v^2 to estimate the probability of selecting each model using the C_p criterion. The value of σ_e^2 was taken to be 1 for all simulations. The results of the simulation are given in Table 2. The values of σ_v^2 considered were 0, 1, 2,

5, 10 and 16 and the values of β' were taken to be $(2, 0, 0, 4, 0)$, $(2, 0, 0, 4, 8)$, $(2, 9, 0, 4, 8)$ and $(2, 9, 6, 4, 8)$ as in Shao (1993). Models were categorized as optimal, category II (correct but not optimal), or category I (incorrect).

The C_p criterion did not perform well for large values of σ_v^2 . For the model $\beta' = (2, 0, 0, 4, 0)$ with $\sigma_v^2 = 1$ the estimated selection probabilities were: optimal model, 0.54; correct model, 0.46; incorrect model, 0. In contrast, when $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.43; correct model, 0.35; incorrect model, 0.22.

The C_p criterion also did not perform well for larger models with large values of σ_v^2 . The C_p criterion however

Table 2
Probabilities of Model Selection Before Transformation

$\beta = (2, 0, 0, 4, 0)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0.62	0.54	0.49	0.46	0.45	0.43
0, 2, 3	II	0.11	0.09	0.09	0.10	0.07	0.06
0, 1, 3	II	0.09	0.14	0.19	0.17	0.15	0.12
0, 3, 4	II	0.09	0.13	0.13	0.14	0.11	0.10
0, 1, 2, 3	II	0.03	0.05	0.06	0.05	0.04	0.04
0, 1, 3, 4	II	0.02	0.03	0.02	0.02	0.02	0.01
0, 2, 3, 4	II	0.02	0.01	0.02	0.02	0.01	0.02
0, 1, 2, 3, 4	II	0.02	0.01	0.00	0.00	0.01	0.00
0, 1	I	0.00	0.00	0.00	0.01	0.07	0.09
0, 2	I	0.00	0.00	0.00	0.01	0.03	0.05
0, 4	I	0.00	0.00	0.00	0.00	0.01	0.04
0, 1, 2	I	0.00	0.00	0.00	0.01	0.01	0.01
0, 1, 4	I	0.00	0.00	0.00	0.01	0.02	0.03
0, 1, 2, 4	I	0.00	0.00	0.00	0.00	0.00	0.00
$\beta = (2, 0, 0, 4, 8)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0.72	0.67	0.63	0.61	0.58	0.49
0, 2, 3, 4	II	0.12	0.12	0.14	0.14	0.11	0.09
0, 1, 3, 4	II	0.12	0.16	0.18	0.14	0.12	0.11
0, 1, 2, 3, 4	II	0.04	0.05	0.05	0.05	0.04	0.04
0, 4	I	0.00	0.00	0.00	0.00	0.01	0.06
0, 1, 4	I	0.00	0.00	0.00	0.02	0.05	0.10
0, 2, 4	I	0.00	0.00	0.00	0.03	0.07	0.10
0, 1, 2, 4	I	0.00	0.00	0.00	0.00	0.01	0.01
$\beta = (2, 9, 0, 4, 8)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0.83	0.78	0.75	0.63	0.39	0.25
0, 1, 2, 3, 4	II	0.17	0.20	0.18	0.13	0.09	0.07
0, 3, 4	I	0.00	0.01	0.03	0.13	0.29	0.35
0, 1, 4	I	0.00	0.00	0.00	0.03	0.11	0.15
0, 2, 3, 4	I	0.00	0.01	0.03	0.07	0.06	0.09
0, 2, 4	I	0.00	0.00	0.00	0.00	0.02	0.05
0, 1, 2, 4	I	0.00	0.00	0.00	0.02	0.04	0.04
$\beta = (2, 9, 6, 4, 8)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	0.98	0.90	0.60	0.29	0.11
0, 2, 3, 4	I	0.00	0.02	0.07	0.24	0.32	0.28
0, 1, 3, 4	I	0.00	0.00	0.02	0.11	0.18	0.23
0, 1, 2, 4	I	0.00	0.00	0.01	0.06	0.13	0.17
0, 3, 4	I	0.00	0.00	0.00	0.00	0.03	0.09
0, 2, 4	I	0.00	0.00	0.00	0.00	0.03	0.10
0, 1, 4	I	0.00	0.00	0.00	0.00	0.01	0.03
0, 1, 3	I	0.00	0.00	0.00	0.00	0.00	0.00

did very well for large models with small values of σ_v^2 . For the full model $\beta' = (2, 9, 6, 4, 8)$ with $\sigma_v^2 = 1$, the estimated selection probabilities were: optimal model, 0.98; correct model, 0.02; incorrect model, 0. In contrast, when $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.11; incorrect model, 0.89. Note that in this scenario there are no correct models other than the optimal model.

In summary, when the C_p criterion is applied to data following the nested error regression model:

1. For any particular model, the estimated probability of selecting the *optimal* model decreases as σ_v^2 increases.
2. For any particular model, the estimated probability of selecting an *incorrect* model increases as σ_v^2 increases.
3. As the number of variables included in the model increases and σ_v^2 increases, the estimated probability of selecting the *optimal* model decreases.
4. As the number of variables included in the model increases and σ_v^2 increases, the estimated probability of selecting an *incorrect* model increases.

The data were then used to estimate the probability of selecting each model using the C_p criterion under the transformation for ρ known. The results of the simulation are given in Table 3. For the model $\beta' = (2, 0, 0, 4, 0)$ with $\sigma_v^2 = 0$ (standard regression model) the estimated selection probabilities were: optimal model, 0.62; correct model, 0.38; incorrect model, 0 (Table 2). Similarly, under the transformation for ρ known with $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.60; correct model, 0.40; incorrect model, 0 (Table 3). For the full model $\beta' = (2, 9, 6, 4, 8)$, the estimated probability of selecting the optimal model was 1 for both the standard regression model (Table 2, $\sigma_v^2 = 0$) and under the transformation for ρ known for all values of σ_v^2 considered (Table 3).

In practice, ρ is unknown and must be estimated from the data. The transformation for ρ unknown is therefore more helpful for practitioners. The results for the transformation with ρ unknown are displayed in Table 4. When ρ was estimated, there was only a small decrease in the estimated probability of selecting the optimal model or a correct model. The largest decrease in the estimated probability of selecting the optimal model was 0.03 for the model with $\beta' = (2, 0, 4, 0)$ and $\sigma_v^2 = 1$, 0.61 for ρ known (Table 3) compared to 0.58 for ρ unknown (Table 4).

Table 3
Probabilities of Model Selection After Transformation, ρ Known

$\beta = (2, 0, 0, 4, 0)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0.61	0.60	0.61	0.61	0.60
0, 3, 4	II	0.11	0.10	0.11	0.11	0.11
0, 2, 3	II	0.10	0.11	0.11	0.10	0.11
0, 1, 3	II	0.09	0.10	0.08	0.09	0.09
0, 1, 2, 3	II	0.04	0.04	0.04	0.04	0.04
0, 1, 3, 4	II	0.03	0.03	0.03	0.02	0.02
0, 2, 3, 4	II	0.02	0.02	0.02	0.02	0.02
0, 1, 2, 3, 4	II	0.01	0.01	0.01	0.01	0.01
$\beta = (2, 0, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0.71	0.71	0.73	0.72	0.71
0, 2, 3, 4	II	0.13	0.12	0.11	0.12	0.13
0, 1, 3, 4	II	0.11	0.12	0.10	0.11	0.11
0, 1, 2, 3, 4	II	0.05	0.05	0.05	0.05	0.05
$\beta = (2, 9, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0.82	0.83	0.83	0.82	0.83
0, 1, 2, 3, 4	II	0.18	0.17	0.17	0.18	0.17
$\beta = (2, 9, 6, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	1.00	1.00	1.00	1.00

Table 4
Probabilities of Model Selection After Transformation, ρ Unknown

		$\beta = (2, 0, 0, 4, 0)'$				
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0.58	0.59	0.60	0.61	0.60
0, 3, 4	II	0.11	0.10	0.11	0.10	0.10
0, 2, 3	II	0.11	0.10	0.11	0.11	0.11
0, 1, 3	II	0.08	0.09	0.10	0.09	0.09
0, 1, 2, 3	II	0.04	0.04	0.03	0.04	0.04
0, 1, 3, 4	II	0.03	0.03	0.02	0.02	0.02
0, 2, 3, 4	II	0.03	0.03	0.02	0.02	0.03
0, 1, 2, 3, 4	II	0.02	0.02	0.01	0.01	0.01
		$\beta = (2, 0, 0, 4, 8)'$				
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0.70	0.70	0.70	0.71	0.70
0, 2, 3, 4	II	0.13	0.14	0.13	0.13	0.13
0, 1, 3, 4	II	0.13	0.11	0.12	0.11	0.12
0, 1, 2, 3, 4	II	0.04	0.05	0.05	0.05	0.05
		$\beta = (2, 9, 0, 4, 8)'$				
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0.82	0.82	0.81	0.83	0.83
0, 1, 2, 3, 4	II	0.18	0.18	0.19	0.17	0.17
		$\beta = (2, 9, 6, 4, 8)'$				
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	1.00	1.00	1.00	1.00

Based on our simulation results, when the C_p criterion is applied to data following the nested error regression model:

1. Under both transformations (ρ known and ρ unknown), the estimated probability of selecting an *incorrect* model was 0.
2. Under the transformation for ρ known, the probability of selecting the *optimal* model was similar to that of the standard regression model.
3. When ρ was estimated, there was only a small decrease in the estimated probability of selecting the optimal model or a correct model.
4. Under both transformations (ρ known and ρ estimated), the C_p criterion performed well, even for larger models with large values of σ_v^2 .
5. The performance of the C_p criterion for the nested error regression model resembles that of the C_p criterion for the standard regression model.

In summary, the C_p criterion does not perform well under the nested error regression model when σ_v^2 is large. When the transformation for ρ unknown (or ρ known) is applied, the model then becomes a standard regression model and the C_p statistic performs accordingly.

Acknowledgements

The research was supported in part by a grant from the Gallup Organization.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Fuller, W.A., and Battese, G.E. (1973). Transformations for estimation of linear models with nested error structures. *Journal of the American Statistical Association*, 68, 626-632.

Gunst, G.F., and Mason, R.L. (1980). *Regression Analysis and Its Application*. New York: Marcel Dekker.

Henderson, C.R. (1953). Estimation of variance and variance components. *Biometrics*, 9, 226-252.

Mallows, C.L. (1973). Some comments on C_p . *Technometrics*, 15, 661-675.

Rao, J.N.K., Sutradhar, B.C. and Yue, K. (1993). Generalized least squares F test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association*, 88, 1388-1391.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

Wu, C.F.J., Holt, D. and Holmes, D.J. (1988). The effect of two-stage sampling on the F Statistic. *Journal of the American Statistical Association*, 83, 150-159.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 20, No. 3, 2004

The Twelfth Morris Hansen Lecture Simple Response Variance: Then and Now Paul P. Biemer	417
Discussion	
Robert M. Groves	441
Keith Rust.....	445
List-based Web Surveys: Quality, Timeliness, and Nonresponse in the Steps of the Participation Flow Monica Pratesi, Katja Lozar Manfreda, Silvia Biffignandi, and Vasja Vehovar	451
The Impact of Coding Error on Time Use Surveys Estimates Patrick Sturgis.....	467
On the Distribution of Random Effects in a Population-based Multi-stage Cluster Sample Survey Obi C. Ukoumunne, Martin C. Gulliford, and Susan Chinn	481
Estimating Marginal Cohort Work Life Expectancies from Cross-sectional Survey Data Markku M. Nurminen, Christopher R. Heathcote, and Brett A. Davis	495
Missing the Mark? Imputation Bias in the Current Population Survey's State Income and Health Insurance Coverage Estimates Michael Davern, Lynn A. Blewett, Boris Bershadsky, and Noreen Arnold	519
Does Voice Matter? An Interactive Voice Response (IVR) Experiment Mick P. Couper, Eleanor Singer, and Roger Tourangeau.....	551
In Other Journals.....	571

Volume 20, No. 4, 2004

Revisions to Official Data on U.S. GNP: A Multivariate Assessment of Different Vintages Kerry D. Patterson and S.M. Heravi.....	573
Discussion	
Dennis Trewin.....	603
Peter van de Ven and George van Leeuwen.....	607
Don M. Eggington.....	615
Robin Lynch and Craig Richardson.....	623
Rejoinder	
Kerry D. Patterson and S.M. Heravi.....	631
The Best Approach to Domain Estimation Precludes Borrowing Strength Victor Estevao and Carl-Erik Särndal.....	645
Perceptions of Disability: The Effect of Self- and Proxy Response Sunghee Lee, Nancy A. Mathiowetz, and Roger Tourangeau.....	671
Maintaining Race and Ethnicity Trend Lines in U.S. Government Surveys Elizabeth Greenberg, Jon Cohen, and Dan Skidmore.....	687
Confidence Intervals for Proportions Estimated from Complex Sample Designs Alistair Gray, Stephen Haslett, and Geoffrey Kuzmich.....	705
Editorial Collaborators.....	725
Index to Volume 20, 2004.....	729

Volume 21, No. 1, 2005

Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model Hui Zheng and Roderick J.A. Little.....	1
The Accuracy of Estimators of Number of Signatories to a Petition Based on a Sample Duncan I. Hedderley and Stephen J. Haslett.....	21
A Two-stage Nonparametric Sample Survey Approach for Testing the Association of Degree of Rurality with Health Services Utilization John S. Preisser, Cicely E. Mitchell, James M. Powers, Thomas A. Arcury, and Wilbert M. Gesler.....	39
Improving Comparability of Existing Data by Response Conversion Stef van Buuren, Sophie Eyres, Alan Tennant, and Marijke Hopman-Rock.....	53
The Nature of Nonresponse in a Medicaid Survey: Causes and Consequences Patricia M. Gallagher, Floyd Jackson Fowler, Jr., and Vickie L. Stringfellow.....	73
Telephone, Internet and Paper Data Collection Modes for the Census 2000 Short Form Sid J. Schneider, David Cantor, Lawrence Malakhoff, Carlos Arieira, Paul Segel, Khaan-Luu Nguyen, and Jennifer Guarino Tancreto.....	89
The Productivity of the Three-step Test-interview (TSTI) Compared to an Expert Review of a Self-administered Questionnaire on Alcohol Consumption Harrie Jansen and Tony Hak.....	103
Underpinning the E-Business Framework. Defining E-Business Concepts and Classifying E-Business Indicators Xander J. de Graaf and Robin H. Muurling.....	121
In Other Journals.....	137

Volume 33, No. 1, March/mars 2005, 1-148

Douglas P. WIENS Editor's report/Rapport du rédacteur en chef	1
Grace Y. YI & Mary E. THOMPSON Marginal and association regression models for longitudinal binary data with drop-outs: a likelihood-based approach	3
Denis BOSQ Estimation suroptimale de la densité par projection	21
John BRAUN, Thierry DUCHESNE & James E. STAFFORD Local likelihood density estimation for interval censored data	39
Zhigang ZHANG, Liuquan SUN, Xingqiu ZHAO & Jianguo SUN Regression analysis of interval-censored failure time data with linear transformation models	61
Alain G. VANDAL, Robert GENTLEMAN & Xuecheng LIU Constrained estimation and likelihood intervals for censored data	71
Jianguo SUN & Liqun SUN Semiparametric linear transformation models for current status data	85
Alexandre X. CARVALHO & Martin A. TANNER Modeling nonlinear time series with local mixtures of generalized linear models	97
Mayer ALVO & Paul CABILIO General scores statistics on ranks in the analysis of unbalanced designs	115
Sudhir R. PAUL & Xing JIANG Testing the homogeneity of several two-parameter populations	131
Acknowledgement of referees' services/Remerciements aux membres des jurys	145
Forthcoming papers/Articles à paraître	146
Volume 33 (2005): Subscription rates/Frais d'abonnement	147

Some of the β_k may be zero and thus various combinations of variables were chosen from $(x_0, x_1, x_2, x_3, x_4)$ to be the predictors used to generate data coming from a nested error regression model. There are $2^p - 1 = 31$ possible models. Each model will be denoted by a subset of $(0, 1, 2, 3, 4)$ that contains the indices of the variables x_i in the model.

Data were generated using 1,000 simulations for several values of σ_v^2 to estimate the probability of selecting each model using the C_p criterion. The value of σ_e^2 was taken to be 1 for all simulations. The results of the simulation are given in Table 2. The values of σ_v^2 considered were 0, 1, 2,

5, 10 and 16 and the values of β' were taken to be $(2, 0, 0, 4, 0)$, $(2, 0, 0, 0, 4, 8)$, $(2, 9, 0, 4, 8)$ and $(2, 9, 6, 4, 8)$ as in Shao (1993). Models were categorized as optimal, category II (correct but not optimal), or category I (incorrect).

The C_p criterion did not perform well for large values of σ_v^2 . For the model $\beta' = (2, 0, 0, 4, 0)$ with $\sigma_v^2 = 1$ the estimated selection probabilities were: optimal model, 0.54; correct model, 0.46; incorrect model, 0. In contrast, when $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.43; correct model, 0.35; incorrect model, 0.22.

The C_p criterion also did not perform well for larger models with large values of σ_v^2 . The C_p criterion however

Table 2
Probabilities of Model Selection Before Transformation

$\beta = (2, 0, 0, 4, 0)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0.62	0.54	0.49	0.46	0.45	0.43
0, 2, 3	II	0.11	0.09	0.09	0.10	0.07	0.06
0, 1, 3	II	0.09	0.14	0.19	0.17	0.15	0.12
0, 3, 4	II	0.09	0.13	0.13	0.14	0.11	0.10
0, 1, 2, 3	II	0.03	0.05	0.06	0.05	0.04	0.04
0, 1, 3, 4	II	0.02	0.03	0.02	0.02	0.02	0.01
0, 2, 3, 4	II	0.02	0.01	0.02	0.02	0.01	0.02
0, 1, 2, 3, 4	II	0.02	0.01	0.00	0.00	0.01	0.00
0, 1	I	0.00	0.00	0.00	0.01	0.07	0.09
0, 2	I	0.00	0.00	0.00	0.01	0.03	0.05
0, 4	I	0.00	0.00	0.00	0.00	0.01	0.04
0, 1, 2	I	0.00	0.00	0.00	0.01	0.01	0.01
0, 1, 4	I	0.00	0.00	0.00	0.01	0.02	0.03
0, 1, 2, 4	I	0.00	0.00	0.00	0.00	0.00	0.00
$\beta = (2, 0, 0, 4, 8)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0.72	0.67	0.63	0.61	0.58	0.49
0, 2, 3, 4	II	0.12	0.12	0.14	0.14	0.11	0.09
0, 1, 3, 4	II	0.12	0.16	0.18	0.14	0.12	0.11
0, 1, 2, 3, 4	II	0.04	0.05	0.05	0.05	0.04	0.04
0, 4	I	0.00	0.00	0.00	0.00	0.01	0.06
0, 1, 4	I	0.00	0.00	0.00	0.02	0.05	0.10
0, 2, 4	I	0.00	0.00	0.00	0.03	0.07	0.10
0, 1, 2, 4	I	0.00	0.00	0.00	0.00	0.01	0.01
$\beta = (2, 9, 0, 4, 8)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0.83	0.78	0.75	0.63	0.39	0.25
0, 1, 2, 3, 4	II	0.17	0.20	0.18	0.13	0.09	0.07
0, 3, 4	I	0.00	0.01	0.03	0.13	0.29	0.35
0, 1, 4	I	0.00	0.00	0.00	0.03	0.11	0.15
0, 2, 3, 4	I	0.00	0.01	0.03	0.07	0.06	0.09
0, 2, 4	I	0.00	0.00	0.00	0.00	0.02	0.05
0, 1, 2, 4	I	0.00	0.00	0.00	0.02	0.04	0.04
$\beta = (2, 9, 6, 4, 8)'$							
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	0.98	0.90	0.60	0.29	0.11
0, 2, 3, 4	I	0.00	0.02	0.07	0.24	0.32	0.28
0, 1, 3, 4	I	0.00	0.00	0.02	0.11	0.18	0.23
0, 1, 2, 4	I	0.00	0.00	0.01	0.06	0.13	0.17
0, 3, 4	I	0.00	0.00	0.00	0.00	0.03	0.09
0, 2, 4	I	0.00	0.00	0.00	0.00	0.03	0.10
0, 1, 4	I	0.00	0.00	0.00	0.00	0.01	0.03
0, 1, 3	I	0.00	0.00	0.00	0.00	0.00	0.00

did very well for large models with small values of σ_v^2 . For the full model $\beta' = (2, 9, 6, 4, 8)$ with $\sigma_v^2 = 1$, the estimated selection probabilities were: optimal model, 0.98; correct model, 0.02; incorrect model, 0. In contrast, when $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.11; incorrect model, 0.89. Note that in this scenario there are no correct models other than the optimal model.

In summary, when the C_p criterion is applied to data following the nested error regression model:

1. For any particular model, the estimated probability of selecting the *optimal* model decreases as σ_v^2 increases.
2. For any particular model, the estimated probability of selecting an *incorrect* model increases as σ_v^2 increases.
3. As the number of variables included in the model increases and σ_v^2 increases, the estimated probability of selecting the *optimal* model decreases.
4. As the number of variables included in the model increases and σ_v^2 increases, the estimated probability of selecting an *incorrect* model increases.

The data were then used to estimate the probability of selecting each model using the C_p criterion under the transformation for ρ known. The results of the simulation are given in Table 3. For the model $\beta' = (2, 0, 0, 4, 0)$ with $\sigma_v^2 = 0$ (standard regression model) the estimated selection probabilities were: optimal model, 0.62; correct model, 0.38; incorrect model, 0 (Table 2). Similarly, under the transformation for ρ known with $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.60; correct model, 0.40; incorrect model, 0 (Table 3). For the full model $\beta' = (2, 9, 6, 4, 8)$, the estimated probability of selecting the optimal model was 1 for both the standard regression model (Table 2, $\sigma_v^2 = 0$) and under the transformation for ρ known for all values of σ_v^2 considered (Table 3).

In practice, ρ is unknown and must be estimated from the data. The transformation for ρ unknown is therefore more helpful for practitioners. The results for the transformation with ρ unknown are displayed in Table 4. When ρ was estimated, there was only a small decrease in the estimated probability of selecting the optimal model or a correct model. The largest decrease in the estimated probability of selecting the optimal model was 0.03 for the model with $\beta' = (2, 0, 4, 0)$ and $\sigma_v^2 = 1$, 0.61 for ρ known (Table 3) compared to 0.58 for ρ unknown (Table 4).

Table 3
Probabilities of Model Selection After Transformation, ρ Known

$\beta = (2, 0, 0, 4, 0)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0.61	0.60	0.61	0.61	0.60
0, 3, 4	II	0.11	0.10	0.11	0.11	0.11
0, 2, 3	II	0.10	0.11	0.11	0.10	0.11
0, 1, 3	II	0.09	0.10	0.08	0.09	0.09
0, 1, 2, 3	II	0.04	0.04	0.04	0.04	0.04
0, 1, 3, 4	II	0.03	0.03	0.03	0.02	0.02
0, 2, 3, 4	II	0.02	0.02	0.02	0.02	0.02
0, 1, 2, 3, 4	II	0.01	0.01	0.01	0.01	0.01
$\beta = (2, 0, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0.71	0.71	0.73	0.72	0.71
0, 2, 3, 4	II	0.13	0.12	0.11	0.12	0.13
0, 1, 3, 4	II	0.11	0.12	0.10	0.11	0.11
0, 1, 2, 3, 4	II	0.05	0.05	0.05	0.05	0.05
$\beta = (2, 9, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0.82	0.83	0.83	0.82	0.83
0, 1, 2, 3, 4	II	0.18	0.17	0.17	0.18	0.17
$\beta = (2, 9, 6, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	1.00	1.00	1.00	1.00

Table 4
Probabilities of Model Selection After Transformation, ρ Unknown

$\beta = (2, 0, 0, 4, 0)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0.58	0.59	0.60	0.61	0.60
0, 3, 4	II	0.11	0.10	0.11	0.10	0.10
0, 2, 3	II	0.11	0.10	0.11	0.11	0.11
0, 1, 3	II	0.08	0.09	0.10	0.09	0.09
0, 1, 2, 3	II	0.04	0.04	0.03	0.04	0.04
0, 1, 3, 4	II	0.03	0.03	0.02	0.02	0.02
0, 2, 3, 4	II	0.03	0.03	0.02	0.02	0.03
0, 1, 2, 3, 4	II	0.02	0.02	0.01	0.01	0.01
$\beta = (2, 0, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0.70	0.70	0.70	0.71	0.70
0, 2, 3, 4	II	0.13	0.14	0.13	0.13	0.13
0, 1, 3, 4	II	0.13	0.11	0.12	0.11	0.12
0, 1, 2, 3, 4	II	0.04	0.05	0.05	0.05	0.05
$\beta = (2, 9, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0.82	0.82	0.81	0.83	0.83
0, 1, 2, 3, 4	II	0.18	0.18	0.19	0.17	0.17
$\beta = (2, 9, 6, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	1.00	1.00	1.00	1.00

Based on our simulation results, when the C_p criterion is applied to data following the nested error regression model:

1. Under both transformations (ρ known and ρ unknown), the estimated probability of selecting an *incorrect* model was 0.
2. Under the transformation for ρ known, the probability of selecting the *optimal* model was similar to that of the standard regression model.
3. When ρ was estimated, there was only a small decrease in the estimated probability of selecting the optimal model or a correct model.
4. Under both transformations (ρ known and ρ estimated), the C_p criterion performed well, even for larger models with large values of σ_v^2 .
5. The performance of the C_p criterion for the nested error regression model resembles that of the C_p criterion for the standard regression model.

In summary, the C_p criterion does not perform well under the nested error regression model when σ_v^2 is large. When the transformation for ρ unknown (or ρ known) is applied, the model then becomes a standard regression model and the C_p statistic performs accordingly.

Acknowledgements

The research was supported in part by a grant from the Gallup Organization.

References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- Fuller, W.A., and Battese, G.E. (1973). Transformations for estimation of linear models with nested error structures. *Journal of the American Statistical Association*, 68, 626-632.
- Gunst, G.F., and Mason, R.L. (1980). *Regression Analysis and Its Application*. New York: Marcel Dekker.
- Henderson, C.R. (1953). Estimation of variance and variance components. *Biometrics*, 9, 226-252.
- Mallows, C.L. (1973). Some comments on C_p . *Technometrics*, 15, 661-675.
- Rao, J.N.K., Sutradhar, B.C. and Yue, K. (1993). Generalized least squares F test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association*, 88, 1388-1391.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.
- Wu, C.F.J., Holt, D. and Holmes, D.J. (1988). The effect of two-stage sampling on the F Statistic. *Journal of the American Statistical Association*, 83, 150-159.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 20, No. 3, 2004

The Twelfth Morris Hansen Lecture Simple Response Variance: Then and Now Paul P. Biemer	417
Discussion	
Robert M. Groves	441
Keith Rust.....	445
List-based Web Surveys: Quality, Timeliness, and Nonresponse in the Steps of the Participation Flow Monica Pratesi, Katja Lozar Manfreda, Silvia Biffignandi, and Vasja Vehovar	451
The Impact of Coding Error on Time Use Surveys Estimates Patrick Sturgis.....	467
On the Distribution of Random Effects in a Population-based Multi-stage Cluster Sample Survey Obi C. Ukoumunne, Martin C. Gulliford, and Susan Chinn	481
Estimating Marginal Cohort Work Life Expectancies from Cross-sectional Survey Data Markku M. Nurminen, Christopher R. Heathcote, and Brett A. Davis	495
Missing the Mark? Imputation Bias in the Current Population Survey's State Income and Health Insurance Coverage Estimates Michael Davern, Lynn A. Blewett, Boris Bershadsky, and Noreen Arnold	519
Does Voice Matter? An Interactive Voice Response (IVR) Experiment Mick P. Couper, Eleanor Singer, and Roger Tourangeau.....	551
In Other Journals.....	571

Volume 20, No. 4, 2004

Revisions to Official Data on U.S. GNP: A Multivariate Assessment of Different Vintages Kerry D. Patterson and S.M. Heravi.....	573
Discussion	
Dennis Trewin.....	603
Peter van de Ven and George van Leeuwen.....	607
Don M. Eggington.....	615
Robin Lynch and Craig Richardson.....	623
Rejoinder	
Kerry D. Patterson and S.M. Heravi.....	631
The Best Approach to Domain Estimation Precludes Borrowing Strength Victor Estevao and Carl-Erik Särndal.....	645
Perceptions of Disability: The Effect of Self- and Proxy Response Sunghee Lee, Nancy A. Mathiowetz, and Roger Tourangeau.....	671
Maintaining Race and Ethnicity Trend Lines in U.S. Government Surveys Elizabeth Greenberg, Jon Cohen, and Dan Skidmore.....	687
Confidence Intervals for Proportions Estimated from Complex Sample Designs Alistair Gray, Stephen Haslett, and Geoffrey Kuzmich.....	705
Editorial Collaborators	725
Index to Volume 20, 2004	729

Volume 21, No. 1, 2005

Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model Hui Zheng and Roderick J.A. Little	1
The Accuracy of Estimators of Number of Signatories to a Petition Based on a Sample Duncan I. Hedderley and Stephen J. Haslett.....	21
A Two-stage Nonparametric Sample Survey Approach for Testing the Association of Degree of Rurality with Health Services Utilization John S. Preisser, Cicely E. Mitchell, James M. Powers, Thomas A. Arcury, and Wilbert M. Gesler.....	39
Improving Comparability of Existing Data by Response Conversion Stef van Buuren, Sophie Eyres, Alan Tennant, and Marijke Hopman-Rock	53
The Nature of Nonresponse in a Medicaid Survey: Causes and Consequences Patricia M. Gallagher, Floyd Jackson Fowler, Jr., and Vickie L. Stringfellow.....	73
Telephone, Internet and Paper Data Collection Modes for the Census 2000 Short Form Sid J. Schneider, David Cantor, Lawrence Malakhoff, Carlos Arieira, Paul Segel, Khaan-Luu Nguyen, and Jennifer Guarino Tancreto.....	89
The Productivity of the Three-step Test-interview (TSTI) Compared to an Expert Review of a Self-administered Questionnaire on Alcohol Consumption Harrie Jansen and Tony Hak.....	103
Underpinning the E-Business Framework. Defining E-Business Concepts and Classifying E-Business Indicators Xander J. de Graaf and Robin H. Muurling.....	121
In Other Journals.....	137

Volume 33, No. 1, March/mars 2005, 1-148

Douglas P. WIENS Editor's report/Rapport du rédacteur en chef	1
Grace Y. YI & Mary E. THOMPSON Marginal and association regression models for longitudinal binary data with drop-outs: a likelihood-based approach	3
Denis BOSQ Estimation suroptimale de la densité par projection	21
John BRAUN, Thierry DUCHESNE & James E. STAFFORD Local likelihood density estimation for interval censored data	39
Zhigang ZHANG, Liuquan SUN, Xingqiu ZHAO & Jianguo SUN Regression analysis of interval-censored failure time data with linear transformation models	61
Alain G. VANDAL, Robert GENTLEMAN & Xuecheng LIU Constrained estimation and likelihood intervals for censored data	71
Jianguo SUN & Lian SUN Semiparametric linear transformation models for current status data	85
Alexandre X. CARVALHO & Martin A. TANNER Modeling nonlinear time series with local mixtures of generalized linear models	97
Mayer ALVO & Paul CABILIO General scores statistics on ranks in the analysis of unbalanced designs	115
Sudhir R. PAUL & Xing JIANG Testing the homogeneity of several two-parameter populations	131
Acknowledgement of referees' services/Remerciements aux membres des jurys	145
Forthcoming papers/Articles à paraître	146
Volume 33 (2005): Subscription rates/Frais d'abonnement	147

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, N° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

- 1. Présentation
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
- 2. Résumé
 - Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

- 3. Rédaction
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ; o, O, 0; 1,).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
- 4. Figures et tableaux
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).
- 5. Bibliographie
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Douglas P. WIENS	Editor's report/Rapport du rédacteur en chef	1
Grace Y. YI & Mary E. THOMPSON	Marginal and association regression models for longitudinal binary data with drop-outs: a likelihood-based approach	3
Denis BOSQ	Estimation suroptimale de la densité par projection	21
John BRAUN, Thierry DUCHESNE & James E. STAFFORD	Local likelihood density estimation for interval censored data	39
Zhigang ZHANG, Liuguan SUN, Xingqiu ZHAO & Jianguo SUN	Regression analysis of interval-censored failure time data with linear transformation models	61
Alain G. VANDAL, Robert GENTLEMAN & Xuecheng LIU	Constrained estimation and likelihood intervals for censored data	71
Jianguo SUN & Liqun SUN	Semiparametric linear transformation models for current status data	85
Alexandre X. CARVALHO & Martin A. TANNER	Modelling nonlinear time series with local mixtures of generalized linear models	97
Mayer ALVO & Paul CABILLIO	General scores statistics on ranks in the analysis of unbalanced designs	115
Sudhir R. PAUL & Xing JIANG	Testing the homogeneity of several two-parameter populations	131
	Acknowledgement of referees' services/Remerciements aux membres des jurys	145
	Forthcoming papers/Articles à paraître	146
	Volume 33 (2005): Subscription rates/Frais d'abonnement	147

Volume 20, No. 4, 2004

Revisions to Official Data on U.S. GNP: A Multivariate Assessment of Different Vintages	Kerry D. Patterson and S.M. Heravi	573
Discussion		
Dennis Trewin		603
Peter van de Ven and George van Leeuwen		607
Don M. Eggington		615
Robin Lynch and Craig Richardson		623
Rejoinder	Kerry D. Patterson and S.M. Heravi	631
The Best Approach to Domain Estimation Precludes Borrowing Strength	Victor Estevao and Carl-Erik Samsdal	645
Perceptions of Disability: The Effect of Self- and Proxy Response	Sunghee Lee, Nancy A. Mathiowetz, and Roger Tourangeau	671
Maintaining Race and Ethnicity Trend Lines in U.S. Government Surveys	Elizabeth Greenberg, Jon Cohen, and Dan Skidmore	687
Confidence Intervals for Proportions Estimated from Complex Sample Designs	Alistair Gray, Stephen Haslett, and Geoffrey Kuzmich	705
Editorial Collaborators		725
Index to Volume 20, 2004		729

Volume 21, No. 1, 2005

Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model	Hui Zheng and Roderick J.A. Little	1
The Accuracy of Estimators of Number of Signatories to a Petition Based on a Sample	Duncan I. Hedderley and Stephen J. Haslett	21
A Two-stage Nonparametric Sample Survey Approach for Testing the Association of Degree of Rurality with Health Services Utilization	John S. Preisser, Cicely E. Mitchell, James M. Powers, Thomas A. Arcury, and Wilbert M. Gesler	39
Improving Comparability of Existing Data by Response Conversion	Stef van Buuren, Sophie Eyres, Alan Tennant, and Marijke Hopman-Rock	53
The Nature of Nonresponse in a Medical Survey: Causes and Consequences	Patricia M. Gallagher, Floyd Jackson Fowler, Jr., and Vickie L. Stringfellow	73
Telephone, Internet and Paper Data Collection Modes for the Census 2000 Short Form	Sid J. Schneider, David Cantor, Lawrence Malakhoff, Carlos Arteira, Paul Segel, Khanh-Luu Nguyen, and Jennifer Guarino Tancreto	89
The Productivity of the Three-step Test-Interview (TSTI) Compared to an Expert Review of a Self-administered Questionnaire on Alcohol Consumption	Harrie Jansen and Tony Hak	103
Underpinning the E-Business Framework. Defining E-Business Concepts and Classifying E-Business Indicators	Xander J. de Graaf and Robin H. Murling	121
In Other Journals		137

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 20, No. 3, 2004

The Twelfth Morris Hansen Lecture Simple Response Variance: Then and Now	Paul P. Biemer	417
Discussion	Robert M. Groves	441
	Keith Rust	445
List-based Web Surveys: Quality, Timeliness, and Nonresponse in the Steps of the Participation Flow	Monica Pratesi, Katja Lozar Manfreda, Silvia Biffignandi, and Vasja Vehovar	451
The Impact of Coding Error on Time Use Surveys Estimates	Patrick Sturgis	467
On the Distribution of Random Effects in a Population-based Multi-stage Cluster Sample Survey	Obi C. Ukwumune, Martin C. Gulliford, and Susan Chin	481
Estimating Marginal Cohort Work Life Expectancies from Cross-sectional Survey Data	Markku M. Nurminen, Christopher R. Heathcote, and Brett A. Davis	495
Missing the Mark? Imputation Bias in the Current Population Survey's State Income and Health Insurance Coverage Estimates	Michael Davern, Lynn A. Blewett, Boris Bershadsky, and Noreen Arnold	519
Does Voice Matter? An Interactive Voice Response (IVR) Experiment	Mick P. Couper, Eleanor Singer, and Roger Tourangeau	551
In Other Journals		571

D'après les résultats de nos simulations, quand le critère de sélection C_p est appliqué à des données obéissant au modèle de régression erreur emboîtée :

L'étude a été financée partiellement par une bourse de l'organisation Gallup.

Bibliographie

Batesse, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Fuller, W.A., et Batesse, G.E. (1973). Transformations for estimation of linear models with nested error structures. *Journal of the American Statistical Association*, 68, 626-632.

Guns, G.F., et Mason, R.L. (1980). *Regression Analysis and Its Application*. New York : Marcel Dekker.

Henderson, C.R. (1953). Estimation of variance and variance components. *Biometrics*, 9, 226-252.

Mallows, C.T. (1973). Some comments on C_p . *Technometrics*, 15, 661-675.

Rao, J.N.K., Sutradhar, B.C. et Yue, K. (1993). Generalized least squares F test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association*, 88, 1388-1391.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

Wu, C.F.J., Holt, D. et Holmes, D.J. (1988). The effect of two-stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.

En résumé, le critère C_p donne de mauvais résultats sous le modèle de régression à erreur emboîtée quand la valeur de σ_v^2 est grande. Quand on applique la transformation pour p inconnu (ou p connu), le modèle devient un modèle de régression standard et la statistique C_p se comporte en conséquence.

- sous les deux transformations (p connu et p inconnu), la probabilité estimée de sélectionner un modèle *incorrect* est 0;
- sous la transformation pour p connu, la probabilité de sélectionner le modèle *optimal* est semblable à celle du modèle de régression standard;
- quand on doit estimer p , la probabilité estimée de sélectionner le modèle optimal ou un modèle correct ne diminue que légèrement;
- sous les deux transformations (p connu et p estimé), le critère C_p donne de bons résultats, même pour des modèles plus grands avec grande valeur de σ_v^2 ;
- les propriétés du critère C_p pour le modèle de régression à erreur emboîtée ressemblent à celles du critère C_p pour le modèle de régression standard.

Tableau 3
Probabilités de sélection du modèle après transformation, p connu

$\beta = (2, 0, 4, 0)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 3	Optimal	0,61	0,60
0, 3, 4	II	0,11	0,10
0, 2, 3	II	0,10	0,11
0, 1, 3	II	0,09	0,08
0, 1, 2, 3	II	0,04	0,04
0, 1, 3, 4	II	0,03	0,03
0, 2, 3, 4	II	0,02	0,02
0, 1, 2, 3, 4	II	0,02	0,01
$\beta = (2, 0, 0, 4, 8)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 3, 4	Optimal	0,71	0,73
0, 2, 3, 4	II	0,13	0,12
0, 1, 3, 4	II	0,11	0,10
0, 1, 2, 3, 4	II	0,05	0,05
$\beta = (2, 9, 0, 4, 8)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 1, 3, 4	Optimal	0,83	0,83
0, 1, 2, 3, 4	II	0,18	0,17
$\beta = (2, 9, 6, 4, 8)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 1, 2, 3, 4	Optimal	1,00	1,00

Tableau 4
Probabilités de sélection du modèle après transformation, p inconnu

$\beta = (2, 0, 0, 4, 0)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 3	Optimal	0,58	0,59
0, 3, 4	II	0,11	0,10
0, 2, 3	II	0,11	0,10
0, 1, 3	II	0,08	0,09
0, 1, 2, 3	II	0,04	0,04
0, 1, 3, 4	II	0,03	0,03
0, 2, 3, 4	II	0,03	0,03
0, 1, 2, 3, 4	II	0,02	0,02
$\beta = (2, 0, 0, 4, 8)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 3, 4	Optimal	0,70	0,70
0, 2, 3, 4	II	0,13	0,14
0, 1, 3, 4	II	0,13	0,12
0, 1, 2, 3, 4	II	0,04	0,05
$\beta = (2, 9, 0, 4, 8)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 1, 3, 4	Optimal	0,82	0,81
0, 1, 2, 3, 4	II	0,18	0,19
$\beta = (2, 9, 6, 4, 8)^T$			
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$
0, 1, 2, 3, 4	Optimal	1,00	1,00

Tableau 2
Probabilités de sélection du modèle avant transformation

$\beta = (2, 0, 0, 4, 0)'$									
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$		
0, 3	Optimal	0,62	0,54	0,49	0,46	0,45	0,43		
0, 2, 3	II	0,11	0,09	0,09	0,10	0,07	0,06		
0, 1, 3	II	0,09	0,14	0,19	0,17	0,15	0,12		
0, 3, 4	II	0,09	0,13	0,13	0,14	0,11	0,10		
0, 1, 2, 3	II	0,03	0,05	0,06	0,05	0,04	0,04		
0, 1, 3, 4	II	0,02	0,03	0,02	0,02	0,02	0,01		
0, 2, 3, 4	II	0,02	0,01	0,00	0,00	0,01	0,01		
0, 1, 2, 3, 4	II	0,00	0,00	0,00	0,01	0,07	0,09		
0, 1	I	0,00	0,00	0,00	0,01	0,03	0,05		
0, 2	I	0,00	0,00	0,00	0,00	0,02	0,03		
0, 4	I	0,00	0,00	0,00	0,00	0,01	0,04		
0, 1, 2	I	0,00	0,00	0,00	0,01	0,01	0,01		
0, 1, 4	I	0,00	0,00	0,00	0,00	0,07	0,10		
0, 1, 2, 4	I	0,00	0,00	0,00	0,00	0,01	0,01		
$\beta = (2, 0, 0, 4, 8)'$									
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$		
0, 3, 4	Optimal	0,72	0,67	0,63	0,61	0,58	0,49		
0, 2, 3, 4	II	0,12	0,12	0,14	0,14	0,11	0,09		
0, 1, 3, 4	II	0,16	0,18	0,18	0,14	0,12	0,11		
0, 1, 2, 3, 4	II	0,04	0,05	0,05	0,05	0,04	0,04		
0, 4	I	0,00	0,00	0,00	0,00	0,01	0,06		
0, 1, 4	I	0,00	0,00	0,00	0,02	0,05	0,10		
0, 2, 4	I	0,00	0,00	0,00	0,03	0,07	0,10		
0, 1, 2, 4	I	0,00	0,00	0,00	0,00	0,01	0,01		
$\beta = (2, 9, 0, 4, 8)'$									
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$		
0, 1, 2, 3, 4	Optimal	0,83	0,78	0,75	0,63	0,39	0,25		
0, 1, 2, 3, 4	II	0,17	0,20	0,18	0,13	0,09	0,07		
0, 3, 4	I	0,00	0,01	0,03	0,13	0,29	0,35		
0, 1, 4	I	0,00	0,00	0,00	0,03	0,11	0,15		
0, 2, 3, 4	I	0,00	0,01	0,03	0,07	0,06	0,05		
0, 2, 4	I	0,00	0,00	0,00	0,00	0,02	0,05		
0, 1, 2, 4	I	0,00	0,00	0,00	0,02	0,04	0,04		
$\beta = (2, 9, 6, 4, 8)'$									
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$		
0, 1, 2, 3, 4	Optimal	1,00	0,98	0,90	0,60	0,29	0,11		
0, 2, 3, 4	I	0,00	0,02	0,07	0,24	0,32	0,28		
0, 1, 3, 4	I	0,00	0,00	0,02	0,11	0,18	0,23		
0, 1, 2, 4	I	0,00	0,00	0,01	0,06	0,13	0,17		
0, 3, 4	I	0,00	0,00	0,00	0,00	0,03	0,09		
0, 2, 3, 4	I	0,00	0,01	0,03	0,07	0,06	0,05		
0, 2, 4	I	0,00	0,00	0,00	0,00	0,02	0,05		
0, 1, 2, 4	I	0,00	0,00	0,00	0,02	0,04	0,04		

Tableau 1
Données pour la simulation de l'erreur emboîtée

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0,3600	0,5300	1,0600	0,5326	0,0900	0,1800	0,5900	0,1855	0,3200	0,4300	0,2804	0,0998
1,3200	2,5200	5,7400	3,6183	0,0200	0,1600	0,2400	0,1572	0,9800	1,0600	2,8000	2,6013
0,0200	0,0700	0,0700	0,3440	0,1800	0,1100	0,3900	0,2879	0,5600	0,6200	1,4559	1,0182
0,9800	1,0600	2,8900	4,0182	0,8500	1,3300	2,7000	2,6013	0,0100	0,0000	0,1540	0,0700
0,1500	0,2500	0,5000	0,6516	0,3800	0,1800	0,4900	0,2436	0,0300	0,0300	0,1300	0,0400
0,1100	0,3500	0,4000	0,1922	0,3900	0,3800	0,9900	1,3979	0,0200	0,2400	0,0800	0,0107
0,0800	0,1300	0,2800	0,0931	0,4300	0,4600	1,4700	2,0138	0,0600	0,0600	0,1300	0,0400
0,6100	0,8500	0,4900	0,0538	0,5700	1,1600	1,8200	1,9356	0,0300	0,0300	0,1300	0,0400
0,0300	0,0300	0,2300	0,0199	0,1300	0,0300	0,0800	0,1050	0,0200	0,2400	0,0800	0,0107
0,0600	0,1100	0,5000	0,0419	0,0400	0,0500	0,1400	0,2207	0,0400	0,0400	0,1300	0,0400
0,0200	0,2400	0,0800	0,1093	0,1300	0,1800	0,2800	0,0180	0,0700	0,0700	0,1300	0,0400
0,0400	0,0400	0,0800	0,0328	0,2000	0,0700	0,0600	0,0962	0,0100	0,0100	0,0300	0,0100

sélection estimées étaient : modèle optimal, 0,43; modèle correct, 0,35; modèle incorrect, 0,22. Le critère C_p n'a pas donné de bons résultats non plus pour les modèles plus grands avec de grandes valeurs de σ_v^2 . En revanche, il a donné de très bons résultats pour les grands modèles avec de petites valeurs de σ_v^2 . Pour le modèle complet $\beta = (2, 9, 6, 4, 8)$, avec $\sigma_v^2 = 1$, les probabilités de sélection estimées étaient : modèle optimal, 0,98; modèle correct, 0,92; modèle incorrect, 0. Comparativement, pour $\sigma_v^2 = 16$, les probabilités de sélection estimées étaient : modèle optimal, 0,11; modèle incorrect, 0,89. Il convient de souligner que, dans ce scénario, le seul modèle correct est le modèle optimal.

En résumé, quand on applique le critère C_p à des données obéissant au modèle de régression à erreur emboîtée :

1. pour n'importe quel modèle, la probabilité estimée de sélection du modèle *optimal* diminue quand σ_v^2 augmente;
2. pour n'importe quel modèle, la probabilité estimée de sélection d'un modèle *incorrect* augmente quand σ_v^2 augmente;
3. à mesure que le nombre de variables incluses dans le modèle augmente et que σ_v^2 augmente, la probabilité estimée de sélection du modèle *optimal* diminue;
4. à mesure que le nombre de variables incluses dans le modèle augmente et que σ_v^2 augmente, la probabilité estimée de sélection d'un modèle *incorrect* augmente.

Nous avons alors utilisé les données pour estimer la probabilité de sélectionner chaque modèle en utilisant le critère C_p sous la transformation pour un coefficient de corrélation ρ connu. Les résultats de la simulation sont donnés dans le tableau 3. Pour le modèle $\beta' = (2, 0, 0, 4, 0)$ avec $\sigma_v^2 = 0$ (modèle de régression standard), les probabilités de sélection estimées étaient : modèle optimal, 0,62; modèle correct, 0,38; modèle incorrect, 0 (tableau 2). Pareillement, sous la transformation pour ρ connu avec $\sigma_v^2 = 16$, les probabilités de sélection estimées étaient : modèle optimal, 0,60; modèle correct, 0,40; modèle incorrect, 0 (tableau 3). Pour le modèle complet $\beta' = (2, 9, 6, 4, 8)$, la probabilité estimée de sélectionner le modèle optimal était 1 pour le modèle de régression standard (tableau 2, $\sigma_v^2 = 0$), ainsi que sous la transformation avec ρ connu pour toutes les valeurs de σ_v^2 envisagées (tableau 3). En pratique, le coefficient de corrélation ρ est inconnu et doit être estimé d'après les données. Par conséquent, la transformation est plus utile pour les praticiens quand ρ est inconnu. Les résultats de la transformation dans ces conditions sont présentés au tableau 4. Quand nous avons estimé p , la probabilité estimée de sélectionner le modèle optimal ou un modèle correct n'a diminué que légèrement. La diminution la plus importante de la probabilité estimée de sélectionner le modèle optimal était 0,03 pour le modèle avec $\beta' = (2, 0, 4, 0)$ et $\sigma_v^2 = 1$, soit 0,61 pour ρ connu (tableau 3) comparativement à 0,58 pour ρ inconnu (tableau 4).

2. Correction pour les corrélations intradomaines

Comme nous l'avons mentionné à la section précédente, les méthodes classiques de sélection du modèle, telles que l'application du critère C_p , ne conviennent pas, puisqu'elles ne tiennent pas compte des corrélations intraspatiales. Wu, Holt et Holmes (1988), ainsi que Rao, Sutradhar et Yue (1993) ont étudié l'effet des méthodes classiques dans le cas du modèle de régression à erreur emboîtée dans un contexte différent.

Considérons le modèle de régression à erreur emboîtée et posons que $\sigma^2 = \sigma_a^2 + \sigma_e^2$ et que ρ est le coefficient de corrélation intradomaine ordinaire $\rho = \sigma_a^2/\sigma^2$. Comme dans Fuller et Battese (1973) et dans Rao et coll. (1993), transformons le modèle de régression à erreur emboîtée en un modèle de régression standard avec erreur i.i.d.

Soit

$$(4) \quad \alpha_i = 1 - \left[\frac{1 - \rho}{1 + (n_i - 1)\rho} \right]^{-1/2},$$

$$(5) \quad y_{ij}^* = y_{ij} - \alpha_i \bar{y}_{.i},$$

$$(6) \quad x_{ij}^* = x_{ij} - \alpha_i \bar{x}_{.i},$$

où $\bar{y}_{.i} = \sum_{j=1}^{n_i} y_{ij}/n_i$ et $\bar{x}_{.i} = \sum_{j=1}^{n_i} x_{ij}/n_i$. Le modèle transformé devient alors

$$(7) \quad y_{ij}^* = x_{ij}^* \beta + e_{ij}^*,$$

pour $j = 1, \dots, n_i, i = 1, \dots, m$ et les e_{ij}^* sont indépendantes et de même loi $N(0, \sigma_e^2)$. Maintenant, nous pouvons appliquer le critère de sélection du modèle standard C_p aux données transformées.

$$(8) \quad \hat{\rho} = \max \left[0, \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2} \right].$$

Pour obtenir les estimateurs des composantes de la variance, représentons par $\{n_i\}$ les résidus de la régression par les moindres carrés ordinaires de $\{y_{ij}^* - \bar{y}_{.i}\}$ sur $\{x_{ij}^* - \bar{x}_{.i}\}$ sans le terme d'ordonnée à l'origine, où $x_{i,l}^* = \sum_{j=1}^{n_i} x_{ij}^* n_i$ pour $l = 1, \dots, p$. Soit $\{r_{ij}^*\}$ les résidus de la régression par les moindres carrés ordinaires de y_{ij}^* sur $\{x_{ij}^*, \dots, x_{ij}^{(p)}\}$ avec le terme d'ordonnée à l'origine.

Les estimateurs de σ_a^2 et σ_e^2 sont donnés par

$$(9) \quad \hat{\sigma}_e^2 = (n - m - p - 1 - \lambda) \sum_{i=1}^m \sum_{j=1}^{n_i} e_{ij}^2,$$

$$(10) \quad \hat{\sigma}_a^2 = n_a^{-1} \left[\sum_{i=1}^m \sum_{j=1}^{n_i} r_{ij}^2 - (n - p - 1) \hat{\sigma}_e^2 \right],$$

$$(11) \quad n_a = n - tr \left[(X'X)^{-1} \sum_{i=1}^m n_i^2 \bar{x}_{.i} \bar{x}_{.i}' \right]$$

où $\lambda = 0$ si le modèle ne contient pas de terme d'ordonnée à l'origine et $\lambda = 1$ autrement. Nous proposons d'appliquer le critère standard de sélection du modèle C_p à ces observations transformées y_{ij}^* et x_{ij}^* .

3. Une étude par simulation

Nous avons réalisé une étude par simulation pour examiner le comportement du critère de sélection du modèle C_p et des transformations proposées pour le modèle de régression à erreur emboîtée. Nous avons considéré le modèle suivant :

$$(12) \quad y_{ij} = \beta_0 x_{ij0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + v_i + e_{ij}$$

pour $i = 1, \dots, 10, n_i \in \{2, \dots, 5\}, j = 1, \dots, n_i$ et $n = 40$. Les v_i suivent une loi $N(0, \sigma_v^2)$ indépendante des e_{ij} qui suivent une loi $N(0, 1)$. Les données x_{ij} sont tirées d'un exemple donné par Gunst et Mason (1980) et inclus dans Shao (1993) (tableau 1). La valeur de x_{ij0} est 1 pour tous

$i = 1, \dots, 10, j = 1, \dots, n_i$. Comme certains coefficients β_k peuvent être nuls, nous avons choisi, à partir de $(x_{01}, x_{11}, x_{21}, x_{31}, x_{41})$, diverses combinaisons de variables comme prédicteurs pour générer les données provenant d'un modèle de régression à erreur emboîtée. Il existe $2^p - 1 = 31$ modèles possibles. Chacun est dénoté par un sous-ensemble de $\{0, 1, 2, 3, 4\}$ qui contient les indices des variables x_j qui y sont incluses.

Pour générer les données, nous avons exécuté 1 000 simulations pour plusieurs valeurs de σ_a^2 afin d'estimer la probabilité de sélection de chaque modèle au moyen du critère C_p . Nous avons donné la valeur 1 à σ_a^2 pour toutes les simulations. Les résultats des simulations sont présentés au tableau 2. Nous avons considéré les valeurs 0, 1, 2, 5, 10 et 16 pour σ_a^2 et fixé les valeurs de β' à $(2, 0, 0, 4, 8)$, $(2, 0, 0, 4, 8)$, $(2, 0, 0, 4, 8)$ et $(2, 0, 0, 4, 8)$ comme dans Shao (1993). Les modèles ont été répartis en trois catégories, à savoir optimal, catégorie II (correct mais non optimal) ou catégorie I (incorrect).

Le critère C_p a donné de mauvais résultats pour les grandes valeurs de σ_a^2 . Pour le modèle $\beta' = (2, 0, 0, 4, 0)$ avec $\sigma_a^2 = 1$, les probabilités de sélection estimées étaient : modèle optimal, 0.54; modèle correct, 0.46; modèle incorrect, 0. Par contre, pour $\sigma_a^2 = 16$, les probabilités de

Une note sur la statistique C_p sous un modèle de régression à erreur emboîtée

Jane L. Meza et P. Lahiri

Résumé

Les modèles de régression à erreur embouée sont utilisés fréquemment pour l'estimation par petits domaines et les problèmes connexes. Cependant, l'application des critères standard de sélection du modèle de régression aux modèles à erreur embouée donne parfois lieu à des méthodes de sélection du modèle inefficaces. Nous illustrons ce point en examinant les propriétés de la statistique C_p au moyen d'une étude par simulation de Monte Carlo. L'inefficacité de la statistique C_p peut, cependant, être corrigée grâce à une transformation appropriée des données.

Mots clés : Statistiques C_p ; modèle de régression à erreur emboîtée; simulation de Monte Carlo.

1. Introduction

Nous examinons les limites d'un critère de sélection

C_p , quand on l'applique au modèle de régression à erreur emboîtée. La statistique C_p (Mallows 1973) est définie par

$$(1) \quad C^d = \frac{\phi_2}{\phi_2 + 2d} = u + 2d$$

où SCR_p est la somme des carrés des résidus et p est le

d'observations et $\hat{\sigma}^2$ est une estimation de σ^2 . Si le modèle est correct, la valeur de C_p doit être semblable ou inférieure à p . Le critère de sélection du modèle C_p est sensible aux valeurs aberrantes et aux écarts par rapport à l'hypothèse d'erreurs i.i.d. suivant une loi normale. La statistique C_p ne peut, par conséquent, être appliquée directement au modèle de régression à erreur emboîtée, pour lequel la structure de l'erreur n'est pas i.i.d.

lequel la structure de l'erreur n'est pas i.i.d.

Nous proposons une transformation des données qui corrige la corrélation intragrappes et permet d'utiliser le critère standard de sélection du modèle C_p . La méthode que nous présentons ici peut être appliquée pour choisir des covariables dans l'analyse des données d'enquête complexes et aux modèles d'estimation par petits domaines. Par exemple, elle pourrait être utilisée pour sélectionner les covariables dans le modèle de régression à erreur emboîtée utilisé par Battese, Harter et Fuller (1988) pour estimer la superficie (en hectares) des cultures de maïs ou de soja pour douze comtés de l'Iowa. Ces auteurs ont utilisé le modèle

: suivan

$$(2) \quad y_{ij}' = \beta + v_i + e_{ij},$$

(2)

f_{in} est la matrice unitaire de dimensions $n_1 \times n_1$. Puisque les erreurs du modèle à erreur emboîtée ne sont pas i.i.d., nous ne pouvons appliquer les procédures de régression standard. L'étude par simulation décrite à la section 3 montre que le critère C_p donne de mauvais résultats sous le modèle de régression à erreur emboîtée. Les transformations envisagées à la section suivante sont utilisées pour transformer le modèle de régression à erreur emboîtée en un modèle de régression standard à erreurs i.i.d. Appliqué à ces observations transformées, le critère C_p donne de nettement meilleurs résultats.

(3)
$$\gamma = X\beta + \varepsilon$$

exprimé sous la forme matricielle suivante

Le modèle de régression à erreur emboîtée peut être dimension $p+1$ de paramètres inconnus.

pour l'unité $j = 1, \dots, n_j$ dans le comté $i = 1, \dots, m_i$, où n_j est la taille de l'échantillon pour le petit domaine i et la taille totale de l'échantillon est $n = \sum_{i=1}^{m_i} n_j$. Les effets de comté, v_i , suivent une loi $N(0, \sigma_v^2)$ indépendante des erreurs aléatoires e_{ij} , qui suivent une loi $N(0, \sigma_e^2)$. La superficie (en hectares) dans l'unité j du comté i est dénotée y_{ij} et $x_j = (1, x_j^{d_1}, \dots, x_j^{d_p})'$ est un vecteur de dimension $d + 1$ des valeurs des covariables x_1, \dots, x_p pour l'unité j dans le comté i . Le vecteur $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ est un vecteur de

Remerciements

Cette étude a été réalisée pendant que le premier auteur était professeur invité au centre de recherche sur les méthodes d'enquête (ZUMA), à Mannheim, en Allemagne.

Bibliographie

Gabler, S., Häder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lynn, P., Häder, S., Gabler, S. et Laaksonen, S. (2004). Methods for Achieving Equivalence of Samples in Cross-National Surveys. ISER Working Paper 2004-09. Disponible au <http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-09.pdf>.

Lynn, P., et Pisati, M. (2005). Improving the quality of sample design for social surveys in Italy: Lessons from the European Social Survey. À paraître.

Tableau 1
Valeurs d'échantillon de b^* , \hat{b} , c_b^2 , c_w^2 , $b|\hat{b}-b^*|$, $b|\hat{b}-b^*|$, $\text{Corr}(w_{g_j}, b_c)$ et ζ pour 15 enquêtes

Pays	b^*	\hat{b}	c_b^2	c_w^2	$b \hat{b}-b^* $	$b \hat{b}-b^* $	$\text{Corr}(w_{g_j}, b_c)$	ζ
Autriche	AT	6,49	7,08	0,08	0,25	6,15	0,34	0,4549
Belgique	BE	6,56	5,79	0,13	0,00	6,56	0,77	0,7060
Suisse	CH	8,83	9,23	0,12	0,21	8,50	0,40	1,7350
République tchèque	CZ	2,94	2,70	0,24	0,25	2,68	0,24	0,4401
Allemagne	DE	18,85	18,13	0,07	0,11	17,42	0,72	0,8757
Espagne	ES	4,96	5,04	0,17	0,22	4,80	0,15	0,4198
Grande-Bretagne	GB	11,11	12,27	0,08	0,22	10,90	0,21	0,5207
Grèce	GR	5,47	5,86	0,09	0,22	5,25	0,22	0,0807
Hongrie	HU	8,68	8,18	0,06	0,00	8,68	0,00	0,2923
Irlande	IE	12,09	11,18	0,13	0,04	12,05	0,05	3,1054
Israël	IL	11,79	12,82	0,12	0,56	9,27	2,53	0,4401
Italie	IT	10,98	10,87	0,26	0,16	11,80	0,83	1,3018
Norvège	NO	44,09	18,68	1,33	0,01	43,32	0,77	
Pologne (rurale)	PL	10,07	9,45	0,06	0,01	9,88	0,19	
Slovenie	SI	10,76	10,13	0,06	0,00	10,76	0,00	

Parmi les pays où $c_b^2 < c_w^2$, le seul pour lequel $\hat{b} < b^*$ et $\zeta > 1$ est la République tchèque (CZ). Il s'agit aussi du pays pour lequel la valeur de \hat{b} est la plus faible. Quand la taille des grappes est particulièrement faible, l'effet du plan définitif est faible et le choix de l'estimateur de b^* pourrait être moins important.

Dans cinq pays, les unités d'échantillonnage ont été sélectionnées individuellement avec probabilités égales (dans les grappes) à partir de registres de la population (BE, DE, HU, PL, SI). Dans ce cas, l'expression (8) (et par conséquent, (10)) est strictement vérifiée, si bien que nous avons $b > b^*$. Pour trois de ces pays (BE, HU, SI), l'échantillon est sélectionné avec probabilités égales, de sorte que nous observons $\hat{b} = b^*$. Il est clair que \hat{b} est supérieur à b en cas d'échantillonnage avec probabilités égales. Dans le cas de l'Allemagne et de la Pologne, il existe une certaine variation inter (mais non intra grappes) des poids de sondage. Cette variation est modeste en Pologne, et $|\hat{b} - b^*| < |\hat{b} - b|$, mais il n'en est pas ainsi en Allemagne, où l'ancien Allemagne de l'Est a été échantillonnée à un taux nettement plus élevé que l'ancienne Allemagne de l'Ouest.

Le plan d'échantillonnage norvégien est le seul qui donne lieu à une variation importante de la taille des grappes à l'étape de la sélection. L'effet très important de cette variation sur $\hat{b} - b^*$ est nettement visible. De nouveau, il s'agit d'une situation où \hat{b} est probablement préférable à \hat{b} en tant que prédicteur de b^* .

Les plans de l'échantillonnage de l'Irlande et de l'Italie comportent, tous deux, la sélection d'adresses à partir de registres électoraux avec probabilité proportionnelle au nombre d'électeurs, puis la sélection d'un résident au hasard pour chaque adresse sélectionnée. Ce genre de plans ne

- plan d'échantillonnage avec probabilités égales où la taille des grappes varie en raison du plan;
- plan d'échantillonnage avec probabilités égales où la taille des grappes ne varie pas en raison du plan, mais varie vraisemblablement à cause de la non-réponse;
- échantillon fondé sur les adresses, où une personne est sélectionnée à chaque adresse, aucune autre source importante de variation des probabilités de sélection n'existe et la taille des grappes ne varie pas en raison du plan de sondage.

5. Conclusion

2004). Malgré cela, les types de plans d'échantillonnage utilisés sont très variés, avant tout à cause de la diversité des bases de sondage disponibles et des objectifs locaux, tels que le souhait de procéder à une analyse infranationale, qui peut nécessiter une stratification disproportionnée par domaine. Nous utilisons ici les données provenant du premier cycle de l'ESF, dont les travaux sur le terrain ont eu lieu en 2002-2003. Des 22 pays participants, 17 ont utilisé un plan d'échantillonnage avec mise en grappes. De ceux-ci, deux n'avaient pas encore fourni de données-échantillons utilisables au moment de la rédaction de l'article. Au tableau 1, nous présentons les valeurs d'échantillon de b^* , \bar{b} , c_b^2 , c_b^w , c_b^2 , \bar{b} , $|b - b^*|$, $|\bar{b} - b^*|$, $|\text{Corr}(w_{c_j}, b_c)|$, et ζ pour les 15 autres pays. Il convient de souligner que le Royaume-Uni et la Pologne avaient tous deux un plan d'échantillonnage à 2 domaines avec mise en grappes uniquement dans un domaine, à savoir la Grande-Bretagne (c'est-à-dire l'Irlande du Nord non comprise) et les régions moins densément peuplées (c'est-à-dire toutes sauf les 42 plus grandes villes), respectivement. Les chiffres présentés au tableau 1 ont trait uniquement au domaine mis en grappes.

Selon (12), nous devrions observer $\bar{b} > b^*$ quand $c_b^w > c_b^2$. Un plan d'échantillonnage courant pour lequel on peut s'attendre à cette inégalité est un plan où la taille des grappes sélectionnées est constante, si bien que la variation de b_c est limitée à celle causée par les différences de non-réponse et b) les échantillons sont obtenus par sélection d'adresses avec probabilités égales, puis sélection aléatoire subséquent d'une personne par adresse, ce qui entraîne une variation des poids de sondage reflétant la variation de la taille du ménage. En tout, six pays ont adopté un plan de sondage de ce genre (AT, CH, ES, GB, GR, IL). Nous observons effectivement que, pour tous ces pays, $\zeta < 1$ et $\bar{b} > b^*$. En outre, pour cinq de ces pays (AT, CH, ES, GB, GR, h = 1, ..., 5), nous pourrions nous attendre à ce que (10) soit une approximation raisonnable, car la seule variation des poids de sondage est celle imputable à la sélection dans un ménage/une adresse. Pour ces pays, nous nous attendrions à ce que \bar{b} donne de meilleurs résultats que b . En effet, $|\bar{b} - b^*| < |\bar{b} - b^*|$ pour quatre des cinq pays, et $(\sum_{j=1}^c \bar{b}_j^2 / \bar{b}^2 - b^*^2) / (\sum_{j=1}^c \bar{b}_j^2 / \bar{b}^2 - b^*^2) = 0,48$. Le seul pays pour lequel \bar{b} ne représente pas une amélioration est l'Espagne, ce qui était prévisible, puisque b est faible. Les grappes de petite taille sont relativement plus sensibles aux effets de la non-réponse et de la variance d'échantillonnage, ce qui entraîne la violation de (10). En Israël, la stratification disproportionnée selon la région géographique est une autre source de variation des poids de sondage. Puisque cette variation cause aussi une violation de (10), nous ne nous attendrions pas nécessairement à ce que \bar{b} représente une amélioration par rapport à \bar{b} en tant que prédicteur de b^* .

L'Enquête sociale européenne (ESF) est une enquête transnationale qui a été conçue en s'efforçant par tous les moyens d'établir une équivalence fonctionnelle approximative entre les plans d'échantillonnage utilisés par les divers pays participants (Lymn, Häder, Gabler et Laaksonen).

4. Exemple : Enquête Sociale Européenne

Les deux coefficients de variation peuvent être estimés en connaissant le plan d'échantillonnage proposé. À la section suivante, nous étudions la sensibilité de prévisions obtenues de cette façon à l'hypothèse (10) en utilisant des données réelles produites au moyen de divers plans d'échantillonnage avec $\text{Cov}(w_{c_j}, b_c) > 0$.

$$b^* = \bar{b} = \frac{1}{\bar{c}^2} \frac{(1 + c_b^w)}{(1 + c_b^2)} \quad (14)$$

Si l'on s'attend à ce que cette covariance soit faible, il pourrait être approprié de prédire b^* comme suit :

$$\bar{w} = \frac{1}{\bar{c}} \sum_{j=1}^c \frac{m}{\bar{c}_j} w_{c_j} = \frac{1}{\bar{c}} \sum_{j=1}^c \bar{b}_j c_j \bar{w}_c$$
$$\bar{b} = \frac{1}{\bar{c}} \sum_{j=1}^c \frac{m}{\bar{c}_j} b_c = \frac{1}{\bar{c}} \sum_{j=1}^c \bar{b}_j c_j \bar{b}_c$$

où

$$\text{Cov}(w_{c_j}, b_c) = \frac{1}{\bar{c}} \sum_{j=1}^c \sum_{f=1}^c \frac{m}{\bar{c}_j} (b_c^f - \bar{b}_c)(w_{c_j}^f - \bar{w}_c) = \frac{1}{\bar{c}^2} \sum_{j=1}^c \sum_{f=1}^c \frac{m^2}{\bar{c}_j \bar{c}_f} (b_c^f - \bar{b}_c)(w_{c_j}^f - \bar{w}_c) \quad (10)$$
$$= \frac{1}{\bar{c}^2} \sum_{j=1}^c \sum_{f=1}^c \frac{m^2}{\bar{c}_j \bar{c}_f} (b_c^f - \bar{b}_c)(w_{c_j}^f - \bar{w}_c) = \frac{1}{\bar{c}^2} \sum_{j=1}^c \sum_{f=1}^c \frac{m^2}{\bar{c}_j \bar{c}_f} (b_c^f - \bar{b}_c)(w_{c_j}^f - \bar{w}_c)$$

où

L'expression (12) donne à penser qu'on peut prédire b^* en prédisant les grandeurs relatives de c_b^2 et c_b^w . Cependant, ce résultat s'applique à une situation particulière,

3. Incidence sur le plan d'échantillonnage

Si l'expression (10) est vérifiée, alors $\zeta = c_b^2 / c_b^w$.

$$b^* \geq \bar{b} \text{ si, et uniquement si } \zeta = \frac{C^2 \text{Cov}(b_c, b_c w_c^2)}{\sum_{j=1}^c m \bar{b}_j^2 \text{Var}(w_{c_j})} \geq 1. \quad (13)$$

c, alors, d'après (6),

Si la variation des poids n'est soumise à aucune contrainte, mais que $\text{Var}(w_{c_j}) > 0$ pour au moins une grappe

(faible).

cause de la variation d'échantillonnage et de la non-réponse disproportionnée (l'effet de celle-ci, pourrait, naturellement, être considérable si la taille des grappes échantillonnées est

2. Relation entre b^* , \bar{b}^w et \bar{b} sous d'autres hypothèses

Soit

$$\bar{w}^c = \frac{1}{b_c} \sum_{b_c} w_{b_c}^c = \sum_{b_c} w_{b_c}^c b_c^{\text{cl}},$$

$$\text{Cov}(b_c, b_c^w w_c^2) = \frac{1}{C} \sum_{c=1}^C b_c^2 w_c^2 - \frac{C^2}{m} \sum_{c=1}^C b_c^2 w_c^2$$

et

$$\frac{1}{b_c} \sum_{b_c} w_{b_c}^c (w_{b_c}^c - \bar{w}^c)^2 = \sum_{b_c} \frac{b_c^{\text{cl}}}{b_c} w_{b_c}^c (w_{b_c}^c - \bar{w}^c)^2 \Delta c.$$

Alors

$$b^* = \frac{C \cdot \text{Cov}(b_c, b_c^w w_c^2) + \bar{b} \sum_{c=1}^C b_c^2 w_c^2 + \sum_{c=1}^C b_c \cdot \text{Var}(w_{b_c}^c) \cdot \sum_{c=1}^C b_c^2 w_c^2}{\sum_{c=1}^C b_c^2 w_c^2} \quad (6)$$

Si l'expression (1) est vérifiée, alors (6) devient :

$$b^* = \bar{b} \left(\frac{\sum_{c=1}^C \text{Var}(w_{b_c}^c) + \sum_{c=1}^C \frac{w_c^2}{C}}{\sum_{c=1}^C \frac{w_c^2}{C}} \right) \quad (7)$$

Par conséquent, dans ce cas, $b^* = \bar{b}$. Si, en outre, les poids sont égaux dans les grappes, c'est-à-dire :

$$w_{b_c}^c = w_c, \forall c \in C$$

alors $b^* = \bar{b}$.
Si l'expression (8) est vérifiée, mais non (1), alors

$$b^* \geq \bar{b} \text{ si, et uniquement si, } \text{Cov}(b_c, b_c^w w_c^2) \geq 0 \text{ puisque}$$

$$b^* - \bar{b} = \frac{C \cdot \text{Cov}(b_c, b_c^w w_c^2)}{\sum_{c=1}^C b_c^2 w_c^2}.$$

La covariance sera négative uniquement si de petites

tailles de grappes coïncident avec de grands poids moyens dans les grappes et inversement. À la section 4, nous observerons que cette situation ne s'est présentée dans aucun pays lors du premier cycle de l'Enquête sociale européenne. De surcroît, de (3) et (4), nous tirons :

$$b^* = \bar{b}^w = \sum_{c=1}^C (w_c^c b_c^c)^2 \bigg/ \sum_{c=1}^C w_c^2 b_c^c \quad (9)$$

Si nous imposons en outre la contrainte (1), alors nous obtenons le résultat évident $b^* = \bar{b}^w = \bar{b} = b^c \Delta c$.

Lynn et Gabler : Approximations de b^* dans la prévision des effets du plan dus à la mise en grappes

Le résultat donné par (9) s'appliquerait aux enquêtes où la seule variation des probabilités de sélection est celle due à un échantillonnage disproportionné entre domaines pour lesquels il n'y a pas de recoupement de grappes. Un exemple courant serait la stratification disproportionnée selon la région, avec les LUPB correspondant à des zones géographiquement contenues dans les régions.

Un assouplissement pratique de la contrainte imposée à la variation des poids est :

$$b_c^c = b_c \left(\frac{b_c^{\text{cl}}}{b_c} \right) \forall c, \quad (10)$$

Autrement dit, nous permettons aux poids de varier dans les grappes, mais nous contrainçons la distribution des fréquences relatives à être la même dans toutes les grappes, c'est-à-dire que les moyennes et les variances des poids dans les grappes ne dépendent pas des grappes.

Maintenant, (3) se simplifie comme suit :

$$b^* = \sum_{c=1}^C \left(\sum_{b_c=1}^{b_c^{\text{cl}}} w_{b_c}^c b_c^c \right) \left(\frac{m}{\sum_{b_c=1}^{b_c^{\text{cl}}} w_{b_c}^c b_c^c} \right) \bigg/ \sum_{c=1}^C \left(\sum_{b_c=1}^{b_c^{\text{cl}}} w_{b_c}^c b_c^c \right) \left(\sum_{b_c=1}^{b_c^{\text{cl}}} w_{b_c}^c b_c^c \right) \bigg/ m^2 \quad (11)$$

Notons que $\left(\sum_{b_c=1}^{b_c^{\text{cl}}} w_{b_c}^c b_c^c \right)^2 / \sum_{b_c=1}^{b_c^{\text{cl}}} w_{b_c}^c b_c^c = m / (1 + c_c^w)$, où c_c^w est le carré du coefficient de variation, sur l'ensemble des observations, des poids. En outre, $(\sum_{c=1}^C c_c^2) / m^2 = (1 + c_c^2) / C$, où c_c^2 est le carré du coefficient de variation, sur l'ensemble des grappes, des tailles de grappe. Donc, (11) devient :

$$b^* = \frac{m}{m} \cdot \frac{(1 + c_c^w)}{(1 + c_c^2)} = \bar{b} \quad (12)$$

Par conséquent, \bar{b} sous-estimera b^* si $c_c^2 > c_c^w$ et inversément. Plus précisément, si $w_{b_c}^c = w \Delta c$, c et $c_c^2 > 0$, alors $b^* > \bar{b}$. La mesure dans laquelle \bar{b} sous-estimera b^* sera d'autant plus importante que la variation de b_c sera grande.

Il est rare que l'hypothèse (10) soit parfaitement vérifiée, mais ce résultat pourrait être utile dans des situations où la distribution des poids devrait, en principe, être la même dans les diverses grappes. Les échantillons fondés sur les adresses où on sélectionne une personne par adresse pouraient en être un exemple. Si la distribution du nombre de personnes par adresse est approximativement constante dans les LUPB (dans la population), alors la distribution des poids ne varierait d'une grappe à l'autre de l'échantillon qu'à

Approximations de b^* dans la prévision des effets du plan dus à la mise en grappes

Peter Lynn et Siegfried Gabler¹

Résumé

Il est fréquent de se servir de l'expression bien connue de l'effet du plan dû à la mise en grappes élaborée par Kish pour éclairer le processus d'échantillonnage en utilisant une approximation telle que b^* à la place de b . Cependant, si le plan comprend une pondération ou une variation de la taille des grappes, cette approximation peut être médiocre. Dans le présent article, nous discutons de la sensibilité de l'approximation aux écarts par rapport aux hypothèses implicites et proposons une approximation de rechange.

Mots clés : Plan d'échantillonnage complexe; coefficient de corrélation intragrappe; probabilité de sélection; pondération.

1. Présentation d'autres fonctions de la taille des grappes

Kish (1965) a utilisé une expression de l'effet du plan (facteur d'accroissement de la variance) dû à la mise en grappes de l'échantillon, $\text{deff} = 1 + (b - 1) p$, où b est le nombre d'observations dans chaque grappe (unité primaire d'échantillonnage) et p est le coefficient de corrélation intragrappe. Cette expression bien connue est enseignée dans les cours sur la théorie de l'échantillonnage et est utilisée par les statisticiens d'enquête pour concevoir et évaluer les échantillons.

L'expression est vérifiée si la taille des grappes ne varie pas et que l'échantillonnage est fait avec probabilités égales (autopondération). Nous pouvons exprimer ces deux critères formellement par :

$$(1) \quad b^c = b^* \quad \text{ou}$$

$$(2) \quad w_i = w^* \quad \text{ou}$$

où $i = 1, \dots, I$ représente les classes de pondération et w_i , les poids de sondage connexes. Cependant, la plupart des enquêtes s'écartent des conditions (1) et (2). Dans le cas général, c'est-à-dire l'élimination des contrantes (1) et (2), Gabler, Häder et Lahiri (1999) ont montré que, sous un modèle approprié, $\text{deff}_c = 1 + (b^* - 1) p$, où

$$b^* = \frac{\sum_{c=1}^C \sum_{i=1}^I w_i b_{ci}}{\sum_{c=1}^C \sum_{i=1}^I w_i^2 b_{ci}} \quad (3)$$

Il est beaucoup plus facile de prédire b^* que b ou b^* à l'étape de l'élaboration du plan d'échantillonnage, car il suffit de connaître le nombre total d'observations, m et le nombre total de grappes, C .

$$(5) \quad \bar{b} = \sum_{c=1}^C b^c / C = m / C.$$

où b^c est le nombre d'observations dans la grappe c , $b^c = \sum_{i=1}^I b_{ci}$. Cependant, (4) n'est pas plus facile à prédire que (3) à l'étape de la conception du plan d'échantillonnage. Une interprétation plus simple, qui est peut-être adoptée fréquemment pour élaborer les plans d'échantillonnage, est la taille de grappe moyenne non pondérée :

$$(4) \quad \bar{b} = \frac{\sum_{c=1}^C \sum_{i=1}^I w_i^2 b_{ci}}{\sum_{c=1}^C \sum_{i=1}^I w_i^2} \quad \text{ou}$$

étant une forme de taille pondérée moyenne de grappe : Gabler et coll. (1999) ont interprété le b de Kish comme d'échantillonnage, la façon de prédire b^* n'est pas claire, chaque observation, le poids de sondage et la grappe microdonnées d'enquête, à condition de connaître, pour La quantité b^* peut alors être calculée d'après les observations dans la grappe c , $j = 1, \dots, b^c$. souci de cohérence) et w_{cj} est le poids associé à la j^{e} modification la notation utilisée par Gabler et coll. (1999) par pondération i dans la grappe c , $b_i = \sum_{c=1}^C b_{ci}$ (nous avons et b_{ci} est le nombre d'observations dans la classe de

$p(X, Y) = 0,9$, $\sigma_x^2 = 10$ et $n = 250$, où la variance estimée de GREG est un peu plus faible. Toutefois, pour la corrélation la plus faible, la diminution de la variance estimée d'OPT comparativement à HT n'est pas importante. Par ailleurs, GREG ne concurrence pas bien les deux autres estimateurs et cette anomalie est particulièrement prononcée pour la plus grande taille d'échantillon $n = 2500$. Donner à $p(X, Y)$ la valeur 0,7 améliore OPT ainsi que GREG, mais ce dernier est de nouveau, dans ces conditions, inférieur à HT dans la plupart des cas. Enfin,

Tableau 1
Efficacité relative estimée (en pourcentage) d'OPT ($S_{y\text{HT}}^2 / S_{y\text{HT}}^2$) et de GREG ($S_{y\text{HT}}^2 / S_{y\text{HT}}^2$) par rapport à HT, basée sur 25 000 échantillons simulés pour chaque taille d'échantillon

$\sigma_x^2 = 10$		$\sigma_x^2 = 100$		$\sigma_x^2 = 100$		$\sigma_y^2 = 100$		$\sigma_y^2 = 100$	
OPT	GREG	OPT	GREG	OPT	GREG	OPT	GREG	OPT	GREG
$p(X, Y) = 0,5$		$p(X, Y) = 0,7$		$p(X, Y) = 0,9$		$n = 250$		$n = 1000$	
99,1	232,8	97,4	176,8	93,9	179,4	91,4	122,3	98,3	247,1
96,8	756,7	96,8	1455,0	97,8	534,7	96,8	1625,5	96,8	141,9
$n = 2500$		$n = 1000$		$n = 250$		$n = 1000$		$n = 2500$	
89,7	197,6	83,8	101,2	73,6	120,4	64,3	72,9	91,0	227,5
93,8	648,2	91,5	1308,6	93,1	218,6	93,1	673,5	93,8	648,2
$p(X, Y) = 0,9$		$n = 250$		$n = 1000$		$n = 2500$		$n = 1000$	
56,5	76,1	41,2	38,8	27,2	43,4	40,4	41,4	61,8	87,3
44,1	44,1	59,8	335,4	63,6	66,0	74,6	259,8	44,1	44,1
$n = 2500$		$n = 1000$		$n = 250$		$n = 1000$		$n = 2500$	
77,0	237,4	59,8	335,4	63,6	66,0	74,6	259,8	77,0	237,4

Bibliographie

Andersson, P.G. (2001). Improving estimation quality in large sample surveys. Thèse de doctorat, Department of Mathematics, Chalmers University of Technology and Göteborg University.

Andersson, P.G., Nerman, O. et Westhall J. (1995). Auxiliary information in survey sampling. *Technical Report NO 1995:3*, Department of Mathematics, Chalmers University of Technology and Göteborg University.

Deville, J.C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E. (2000). Conditioning on auxiliary variable means in finite population inference. *Australian & New Zealand Journal of Statistics*, 42, 407-421.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

d'échantillonnage aléatoire simple stratifié, notamment si les pentes ne sont pas les mêmes pour les diverses strates et la population non stratifiée. Considérons, par conséquent, une population de taille N , qui est partitionnée en L strates N_1, \dots, N_L . Tirons, à partir de chaque strate, n_h un échantillon aléatoire simple s_h de taille n_h , où $s_1 + \dots + s_L = s$ et $n_1 + \dots + n_L = n$. Par souci de simplicité, supposons aussi que l'information auxiliaire est unidimensionnelle et globale, c'est-à-dire que seul t_x est connu d'avance. Pour GREG, nous avons choisi le modèle de régression linéaire simple homoscédastique (voir Sändal et coll. 1992).

Les expressions résultantes pour HT, OPT et GREG, respectivement sont

$$\begin{aligned} \hat{t}_y &= N \bar{y}^{st} \\ \hat{t}_{y^{opt}} &= N(\bar{y}^{st} + b^{opt}(\bar{x} - \bar{x}^{st})) \\ \hat{t}_{y^{st}} &= N(\bar{y}^{st} + b^{st}(\bar{x} - \bar{x}^{st})), \end{aligned}$$

où $\bar{x} = (1/N) \sum_{h=1}^L N_h \bar{x}_h^{st}$, $\bar{y}^{st} = (1/N) \sum_{h=1}^L N_h \bar{y}_h^{st}$, (analogue à \bar{x}^{st}) et

$$\begin{aligned} b^{opt} &= \frac{\sum_{h=1}^L N_h^2 \left(\frac{1}{N} - \frac{1}{N_h} \right) \sum_{i \in s_h} n_h (x_i - \bar{x}_h)^2}{\sum_{h=1}^L N_h^2 \left(\frac{1}{N} - \frac{1}{N_h} \right) \sum_{i \in s_h} n_h (y_i - \bar{y}_h)^2} \\ b^{st} &= \frac{\sum_{h=1}^L N_h^2 \sum_{i \in s_h} n_h (x_i - \bar{x}^{st})(y_i - \bar{y}^{st})}{\sum_{h=1}^L N_h^2 \sum_{i \in s_h} n_h (x_i - \bar{x}_h^{st})^2}. \end{aligned}$$

Il est facile de voir d'après ces formules que le coefficient de régression optimal est la moyenne des pentes dans les strates et que le coefficient de régression GREG est la pente globale. Si l'écart entre ces pentes est important, la correction GREG devient mauvaise. Ici, nous nous intéressons tout spécialement à la comparaison des qualités de ces estimateurs lorsque le modèle (linéaire) auxiliaire pour GREG échoue. Nous avons donc généré des valeurs de x et de y à partir de variables aléatoires X et Y corrélées suivant des lois log-normales, où $\ln X$ suit une loi normale d'espérance 0 et de variance $\sigma_1^2(N(0, \sigma_1))$, et $\ln Y$ est $N(0, \sigma_2)$. Les variances σ_1^2 et σ_2^2 , et la corrélation entre $\ln X$ et $\ln Y$, peuvent alors être choisis de façon à obtenir X et de leur corrélation $\rho(X, Y)$. Les valeurs générées à partir des lois normales bivariées ont été obtenues au moyen de MATLAB (version 6.0). Nous avons créé de cette façon 12 populations, chacune de taille $N = 10\,000$, y compris quatre combinaisons de variances σ_x^2 et σ_y^2 (10 et 100) et trois valeurs de corrélation $\rho(X, Y)$ (0,5, 0,7 et 0,9). Pour

5.2 Résultats

Permettrons, à titre de référence, la valeur absolue du biais relatif estimé de l'estimateur HT sans biais n'a excédé dans aucun cas $4 \cdot 10^{-4}$. Les valeurs maximales correspondantes pour OPT et GREG étaient $6 \cdot 10^{-3}$, ce qui signifie que nous pouvons nous concentrer sur les ratios des variances estimées pour évaluer les efficacités relatives de HT, OPT et GREG.

Comme le montre l'examen du tableau 1, OPT est supérieur à HT ainsi qu'à GREG (à une exception près :

$$\frac{S_{y^{opt}}^2}{S_{y^{HT}}^2} \text{ et } \frac{S_{y^{HT}}^2}{S_{y^{HT}}^2},$$

où $\bar{t} = (1/K) \sum_{i=1}^K t_i$. Puisque nous nous intéressons principalement aux comparaisons entre OPT et GREG, nous ne présenterons que les résultats des mesures relatives de la variance (ou, de façon équivalente, l'écart-type)

$$S_z^2 = \frac{1}{K} \sum_{i=1}^K (t_i - \bar{t})^2,$$

et la variance estimée

$$\frac{\bar{t}}{\bar{t} - t}$$

biases relatif estimé

estimateur \bar{t} d'un total t pour une série t_1, \dots, t_L sont le En général, les mesures courantes de la qualité d'un échantillon.

calculé les estimateurs HT, OPT et GREG pour chaque 25 000 pour chacun des $12 \times 3 = 36$ cas, et nous avons alors (n_s/N_s) . Le nombre d'échantillons simulés était $K =$ strate 5 ayant la probabilité de sélection la plus grande portionnelle à la taille) avec, par exemple, les objets dans la d'échantillonnage approximativement PPT (probabilité proportionnelle à la taille) avec, par exemple, les objets dans la

$n_1 = \dots = n_s$. Autrement dit, nous avons créé un plan $n = 250, 1\,000$ et $2\,500$, où pour chaque échantillon $n = 250, 1\,000$ et $2\,500$, nous avons tiré des échantillons de taille constituent la strate 1, ainsi de suite. À partir de chaque po-

que les objets ayant les valeurs de y les plus petites et $N_s = 500$. Ces données sont construites de telle façon tailles $N_1 = 4\,000$, $N_2 = 2\,500$, $N_3 = 2\,000$, $N_4 = 1\,000$ croissantes. Le nombre de strates $L = 5$ est partout avec les objets de chaque population en fonction des valeurs de y Maintenant, avant la stratification, nous ordonnons les 38,59.

ces populations, une variance de 10 implique une asymétrie de 9,37 et une variance de 100 donne une asymétrie de

existe aussi des valeurs de x telles que $XR^{\text{opt}}_T X^T$ ne soit pas semi-définie positive, mais la probabilité de l'existence de telles valeurs de x tend vers zéro quand les tailles de la population et de l'échantillon augmentent (et si $\sum_{i=1}^n x_i$ est une définie positive). Une minimisation stricte d'une distance ayant une « composante négative » entraînerait des corrections infiniment grandes. Autant que nous sachions, ce problème ne pose par l'estimateur optimal n'a encore jamais été souligné.

Le moyen le plus simple de trouver une distance qui donne l'estimateur optimal sous forme d'un estimateur par calage consiste à trouver une matrice R^{dist} ayant les mêmes vecteurs propres que R^{opt} , mais où les valeurs propres sont remplacées par leurs valeurs absolues (ce résultat peut être démontré de la même façon que la preuve du lemme qui précède. La distance peut être considérée comme la somme des produits des valeurs propres et des carrés des vecteurs propres. Écrire que les dérivées sont nulles signifie que, dans la proposition, nous trouvons les extrêmes, c'est-à-dire les minima pour les valeurs propres positives et les maxima pour les valeurs propres négatives. Si nous changeons tous les signes négatifs, tous les extrêmes seront des minima).

4. Exemples

Définie positive R^{opt} : Supposons que les objets compris dans U sont sélectionnés indépendamment avec probabilité de sélection π_1, \dots, π_N (échantillonnage de Poisson), ce qui sous-entend une taille d'échantillon aléatoire n , où $E[n] = \sum_{i \in U} \pi_i$. Étant donné l'indépendance des tirages, R^{opt} est diagonale et plus précisément

$$R^{-1}_{\text{opt}} = I_n \left(\frac{\pi_i^2}{1 - \pi_i} \right)_{i \in s}.$$

Semi-définie positive R^{opt} : Supposons que n objets sont tirés selon un plan d'échantillonnage aléatoire simple, c'est-à-dire que chaque objet a une probabilité de sélection $\pi_i = n/N$. Les éléments de R^{opt} sont

$$i = j : \left(\frac{n}{N} \right)^2 \frac{N}{N - n} \quad i \neq j : \left(\frac{n}{N} \right)^2 \frac{n}{N - n}.$$

Cela signifie que R^{opt} est une matrice singulière de rang $n - 1$.

Supposons plutôt (comme dans l'étude par simulation suivante) que U est partitionnée en L strates de tailles N_1, \dots, N_L , à partir desquelles nous tirons des échantillons aléatoires simples indépendants de tailles n_1, \dots, n_L . Les éléments de R^{opt} sont alors

5. Une étude par simulation

5.1 Notation et aperçu

Afin de comparer empiriquement l'estimateur optimal (OPT) à l'estimateur GREG (GREG), et de comparer également ces deux estimateurs à l'estimateur d'Horvitz-Thompson (HT), nous avons réalisé une petite étude par simulation. Aux sections précédentes, nous avons mentionné que OPT est l'estimateur linéaire asymptotiquement le plus efficace et un estimateur par calage. Bien qu'il possède de nombreuses propriétés intéressantes, il peut être inefficace pour des tailles d'échantillon raisonnables. Ici, nous allons montrer, au moyen de certaines situations simulées, que l'estimateur optimal peut aussi représenter une amélioration considérable par rapport à GREG pour des tailles d'échantillon moyennes, quand la population est (délibérément) choisie de façon telle qu'elle soit défavorable pour GREG. Une situation simple, mais non triviale, pour laquelle OPT n'est pas égal à GREG se produit en cas

avec probabilité de 0,04. La deuxième matrice a une valeur propre négative.

Le problème ne disparaît pas nécessairement si N est grand. Considérons, au lieu de la précédente, une population comprenant $N/4$ strates contenant chacune quatre éléments. Supposons que la procédure d'échantillonnage sus-mentionnée soit utilisée indépendamment dans chaque strate. Dans ce cas, R^{opt} sera une matrice constituée des matrices 2×2 susmentionnées sur la diagonale et de zéro ailleurs.

$$R^{\text{opt}} = \begin{pmatrix} -96 & 2 \\ 2 & -96 \end{pmatrix} \quad (13)$$

avec probabilité de 0,96 et

$$R^{\text{opt}} = \begin{pmatrix} 2 & 23/12 \\ 23/12 & 2 \end{pmatrix} \quad (12)$$

0,01. Dans ce cas, c'est-à-dire $\pi_{13} = \pi_{24} = 0,48$ et $\pi_{12} = \pi_{14} = \pi_{23} = \pi_{34} = 0,06$, qu'un échantillon systématique est tiré avec probabilité 0,94 et un échantillon aléatoire simple, avec probabilité de 0,06, est de rang $N - h$.

strate h , $h = 1, \dots, L$ et 0 autrement. Par conséquent, R^{opt} est de rang $N - h$.

où, dans le dernier cas, i et j appartiennent tous deux à la

$$i = j : \left(\frac{n_h}{N_h} \right)^2 \frac{n_h}{N_h - n_h} \quad i \neq j : \left(\frac{n_h}{N_h} \right)^2 \frac{n_h(n_h - 1)}{N_h - N_h},$$

qui est appelée équation de calage. Ceci nous amène à un nature moyen possible d'obtenir l'estimateur GREG, conformément à Deville et Särndal (1992). Supposons que nous cherchions un estimateur y^*w de t_y avec un vecteur w de poids dépendant de l'échantillon $(w_i)_{i \in s}$, qui respecte l'équation de calage correspondante, tout en minimisant la distance entre w et w_0 conformément à la mesure de distance quadratique

où $\mathbf{R} = (w^0 \mathbf{I}^n)^{-1}$. Ceci nous donne

$$(4) \quad \left(\begin{matrix} x \\ 1 \end{matrix} \right)_{-1}^x \left(\begin{matrix} x \\ 1 \end{matrix} \right)_{-1}^x X R X \left(\begin{matrix} x \\ 1 \end{matrix} \right)_{-1}^x + R + {}^0 M = M$$

qui signifie que $w = g$, puisque ici $R = R^{-1}$.

En ce qui concerne l'estimateur optimal, considérons d'abord le vecteur (\hat{f}_y, \hat{f}_x^T) et posons que $\Sigma_{y,x}$ est le vecteur (ligne) des covariances de \hat{f}_y et \hat{f}_x , et que $\Sigma_{x,x}$ est la matrice des covariances de \hat{f}_x . Maintenant, l'estimateur linéaire sans biais (en \hat{f}_y et \hat{f}_x) à variance minimale (voir Montmarini [1987] de t_y est l'estimateur par différence

$$(5) \quad \cdot \left(\underset{\downarrow}{x} 1 - \underset{\downarrow}{x} 1 \right) \underset{\downarrow}{x} x \Sigma \underset{\downarrow}{x} x \Sigma + \underset{\downarrow}{x} 1$$

Puisqu'en pratique, $\sum_{y,x}$ et $\sum_{x,x}$ sont inconnus, posons

$$(9) \quad \begin{aligned} &({}^x J - {}^x I)_{I-X} ({}_I X {}^R (X) ({}_I X {}^R {}_I X) + {}^x J = \\ &({}^x J - {}^x I)_{I-X} {}^x J + {}^0 M {}_I X = {}^{10} J \end{aligned}$$

où $R^{\text{opt}} = ((\pi_{\hat{f}} - \pi_l \pi_f) / (\pi_{\hat{f}} \pi_l \pi_f))_{l, f \in \mathcal{S}}$. Dans un contexte asymptotique, où $n \rightarrow \infty$ et

$$(L) \quad ({}^0w - w)_{I^-} R_I ({}^0w - w)$$

Donc, le fait que l'estimateur GREG soit aussi un estimateur par calage dont la mesure de distance est

et la comparaison de (1) et (6) portent à croire que remplacer \mathbf{R} , par \mathbf{R}^{opt} dans (7) devrait impliquer que nous pliot l'estimateur par la régression optimale que nous décrivons sous forme d'un estimateur par calage. Nous montrons plus loin que cela est vrai.

3. Le résultat principal

Afin de montrer l'existence d'une mesure de distance correspondant à l'estimateur optimal, nous commencerons par énoncer et prouver un résultat dans le cas général.

Lemme : Avec R dénotant une matrice définie positive arbitraire de dimensions $n \times n$,

$$(8) \quad ({}^0M - M) R_T ({}^0M - M)$$

soumise à la contrainte $Xw = t^x$, est minimisée par

$$({}^x I - {}^x I)_{I-} ({}_I X {}_I R {}_I X) {}_I X {}_I R + {}^0 M = M$$

Démonstration : En introduisant le vecteur λ de dimensions $J \times 1$ de multiplicateurs de Lagrange, après dérivation, nous obtenons le système d'équations

$$(6) \quad 0 = \gamma_T X + (w^0 - w) Z$$

En multipliant (9) par XR^{-1} , en utilisant (10) et en isolant λ , nous obtenons, avec $Xw_0 = \hat{f}_x$:

$$(11) \quad \lambda = 2(XR^T X - I)^{-1} (I - I^x).$$

En introduisant cette expression dans (9) et en isolant w nous obtenons finalement

$$w = w_0 + R_{-1} X (X' R_{-1} X)^{-1} X' (I - I_{-1}^x) \downarrow (I - I_{-1}^x) w$$

D'après le lemme, nous avons donc le résultat principal suivant :

Théorème : En posant que R^{opt} est une (semi)-définition de distance de calage optimale, que nous obtenons en permettant que $R = R^{\text{opt}}$ dans (8), l'estimateur par calage deviendra l'estimateur par la régression optimale.

Remarque : Dans certains cas, R^{opt} peut être indéfinie (voir plus loin). La seule chose que nous savons est qu'il s'agit d'un estimateur sans biais d'une matrice de covariance. Si elle n'est pas une semi-définie positive, il

comparatives (voir Anderson et coll. 1995), l'estimateur optimal à la même variance asymptotique que l'estimateur par différence (5). En particulier, il s'ensuit que l'estimateur optimal est asymptotiquement meilleur que l'estimateur GREG habituel (voir Rao 1994, Montanari 2000 et Anderson 2001), c'est-à-dire sa variance asymptotique n'est jamais plus grande et est habituellement plus faible. À la section 5, nous présentons certaines simulations simples montrant que l'estimateur optimal peut être nettement plus efficace que l'estimateur GREG. Cependant, on ne sait rien de l'efficacité des échantillons fins, puisque l'estimateur de la covariance peut converger lentement. La vitesse de convergence est illustrée à la section 5. Notons aussi que, dans certains cas, il existe des estimateurs non linéaires qui sont asymptotiquement encore meilleurs.

où $R^{\text{opt}} = ((\pi_j - \pi_i, \pi_j) / (\pi_j, \pi_i))_{i, j \in \mathcal{S}}$. Dans un contexte asymptotique, où $n \rightarrow \infty$ et $N \rightarrow \infty$, $\sum_{x \in \mathcal{X}} \pi_x$, et $\sum_{x \in \mathcal{X}} \pi_x$ peuvent être considérés comme des composantes de la matrice asymptotique des covariances de $\begin{pmatrix} \sum_{x \in \mathcal{X}} \pi_x \\ \sum_{i, j \in \mathcal{S}} (\pi_j - \pi_i, \pi_j) \end{pmatrix}$. Sous l'hypothèse de convergence de $\sum_{x \in \mathcal{X}} \pi_x$, qui est vérifiée sous des conditions très peu

$$(9) \quad \begin{pmatrix} x & y \\ t & t' \end{pmatrix} = \begin{pmatrix} X & Y \\ R & R' \end{pmatrix} \begin{pmatrix} X & Y \\ R & R' \end{pmatrix}^{-1}$$
$$\hat{f}^{\text{do}} = \sum_{\mathbf{y}} \mathbf{M}^T \mathbf{y} = \sum_{\mathbf{y}} (\mathbf{I} - \mathbf{I}^x) \mathbf{y} = \mathbf{I}^x \mathbf{x}^*$$
$$\hat{f}_y + \sum \hat{f}_{y,x} - \sum \hat{f}_{x,x} = 0. \quad (5)$$

linéaire sans biais (en \hat{t}_y et \hat{t}_x) à variance minimale (voir Montanari 1987) de t_y est l'estimateur par différence

d'abord le vecteur (\hat{t}_y, \hat{t}_T^x) et posons que $\sum_{y,x}$ est le vecteur (ligne) des covariances de \hat{t}_y et \hat{t}_T^x , et que $\sum_{x,x}$ est la matrice des covariances de \hat{t}_x . Maintenant, l'estimateur

qui signifie que $w = g$, puisque ici $R = R^{-1}$.

$$(4) \quad w = w^0 + R_{l-1}^T X R_{l-1} (I_{l-1} - x x^T) x, \quad (I_{l-1} - x x^T) x = 0,$$

Ceci nous donne

où $\mathbf{R} = I^n$.

$$({}^0w - w)R_L({}^0w - w)$$

distance entre w et w_0 conformément à la mesure de distance quadratique

Pour dépendant de l'échelle ($W_{1/2}^i$), qui respecte l'équation de calage correspondante, tout en minimisant la

chérchions un estimateur $y_T w$ de t_y avec un vecteur w de

autre moyen possible d'obtenir l'estimateur GREG, conformément à Deville et Särndal (1992). Supposons que nous

qui est appelée équation de calage. Ceci nous amène à un

Andersson et Thorburn : Une distance de calage optimale menant à un estimateur par la régression optimal

Une distance de calage optimale menant à un estimateur par la régression optimal

Per Gösta Andersson et Daniel Thorburn

Résumé

En échantillonnage, quand on dispose d'information auxiliaire, il est bien connu que l'estimateur (par la régression) optimal « fondé sur le plan de sondage d'un total ou d'une moyenne de population finie est (du moins asymptotiquement) plus efficace que l'estimateur GREC correspondant. Nous illustrerons ce fait au moyen de simulations avec échantillonnage stratifié à partir de populations à distribution asymétrique. Au départ, l'estimateur GREC a été construit au moyen d'un modèle linéaire de superpopulation auxiliaire. Il peut aussi être considéré comme un estimateur par calage, c'est-à-dire un estimateur linéaire pondéré, où les poids obéissent à l'équation de calage et, sous cette contrainte, sont aussi proches que possible des « poids d'Horvitz-Thompson » originaux (d'après une mesure de distance appropriée). Nous montrons que l'estimateur optimal peut aussi être considéré comme un estimateur par calage à cet égard avec une mesure quadratique de distance étroitement liée à celle générant l'estimateur GREC. Nous donnons aussi des exemples simples révélant qu'il n'est pas toujours facile d'obtenir cette nouvelle mesure.

Mots clés : Estimateur d'Horvitz-Thompson; estimateur par la régression; théorie de l'échantillonnage.

1. Notation et notions élémentaires

Considérons une population finie U constituée de N objets étiquetés $1, \dots, N$ avec les valeurs étudiées connexes y_1, \dots, y_N et les vecteur (colonnes) auxiliaires de dimension $t_y = \sum_{i \in U} y_i$, en tirant un échantillon aléatoire s de taille n (fixe ou aléatoire) à partir de U , avec les probabilités de sélection de premier et de deuxième ordre $\pi_i = P(i \in s)$, $\pi_{ij} = P(i, j \in s)$, $j = 1, \dots, N$. Les valeurs étudiées et les vecteurs auxiliaires sont enregistrés pour les objets échantillonnés et nous supposons qu'au moins $t_x = \sum_{i \in U} x_i$ est connu avant que l'échantillon ne soit tiré.

À la section 2, nous discutons de divers estimateurs par la régression, à savoir l'estimateur GREC ordinaire (Särndal, Swensson et Wretman 1992), l'estimateur optimal (Montanari 1987 et Andersson, Nerman et Westhall 1995) et les estimateurs par calage (Deville et Särndal 1992). Il est bien connu que l'estimateur GREC section 3, nous montrons que cela est également vrai pour l'estimateur optimal, mais avec une mesure de distance plus compliquée. Aux deux dernières sections, nous illustrons ceci et l'estimateur optimal, d'abord au moyen d'exemples théoriques, puis de simulations.

Enfin, suivent certains commentaires au sujet de la notation matricielle utilisée dans l'article. En général, la transposée d'une matrice A est notée A^T et, si A est carrée, l'inverse (inverse généralisée) s'écrit $A^{-1}(A^-)$. En

2. Estimateurs par la régression et par calage

Un estimateur simple sans biais de t_y est l'estimateur d'Horvitz-Thompson $\hat{t}_y = \sum_{i \in s} y_i / \pi_i = y^T w_0$. Cependant, on peut obtenir des estimateurs plus efficaces en utilisant l'information auxiliaire, par exemple, l'estimateur GREC assisté par modèle bien connu (voir Särndal et coll. (1992). Par exemple, sous l'hypothèse d'un modèle de régression linéaire homoscedastique de superpopulation, l'estimateur GREC est

$$\hat{t}_{y^r} = y^T w_0 + (y^T R_p R_p^T X^T)^{-1} (t_x - \hat{t}_x) \quad (1)$$

$$= y^T g, \quad \text{où } R_p = w_0 I_n, \hat{t}_x = \sum_{i \in s} x_i / \pi_i \text{ et} \quad (2)$$

$$g = \left(\frac{1}{\pi_i} (1 + x_i^T (X R_p X^T)^{-1} (t_x - \hat{t}_x)) \right)_{i \in s}.$$

Or, l'expression (2) de l'estimateur GREC est intéressante, puisque nous avons aussi l'expression

$$x^T g = t_x, \quad (3)$$

Natural Resources Conservation Service et la Iowa State University. Nous remercions le rédacteur adjoint et les examinateurs de leurs commentaires qui nous ont permis d'améliorer l'article.

Bibliographie

- Bardsley, P., et Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Deville, J.-C., et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Fuller, W.A. (2002). Estimation par regression appliqué à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- Husain, M. (1969). Construction of Regression Weights for Estimation in Sample Surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.
- Rao, J.N.K., et Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods, American Statistical Association*, 57-64.
- Stephan, F.F. (1942). An alternative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.
- Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.
- Tillé, Y. (1999). Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles : Enquêtes à plan d'échantillonnage complexe. *Techniques d'enquête*, 25, 61-71.

Donc, en vertu de (26), (27) et (28),

$$[N\pi^{[x]_{\text{int}}} - 1] = (N\pi_i - 1)[I + (\bar{x}_N - \bar{x}_{\text{HT}}) \sum_{j=1}^{x-1} \mathbf{d}_j^x] + O_p(n^{-2}).$$

En vertu des hypothèses (18), (23) et (25), et en utilisant le fait que $E\{\pi^{[x]_{\text{int}}}\}$ est bornée,

$$\bar{y}_{p\pi} = \bar{y}_{\text{HT}} + (\bar{x}_N - \bar{x}_{\text{HT}}) \bar{\theta} + O_p(n^{-1})$$

$$= \bar{y}_{\text{HT}} + (\bar{x}_N - \bar{x}_{\text{HT}}) \bar{\theta} + O_p(n^{-1}). \quad (29)$$

Si la valeur \mathbf{x}_i est un élément de $\mathbf{X}_{\bar{y}}$, l'estimateur SCP de $\text{Var}(\sum_{i=1}^n \pi_i^{-1} \mathbf{x}_i) = 0$, et si $\mathbf{M}_{\bar{y}} = \sum_{\bar{y}} \mathbf{x}_i$, l'estimateur SCP de la moyenne de population du vecteur $\mathbf{q}_i = (1, \mathbf{x}_i)$ satisfait

$$\bar{\mathbf{q}}_{p\pi} = N^{-1} \sum_{i=1}^n \pi_i^{-1} \mathbf{q}_i = (1, \bar{x}_N) + O_p(n^{-1}), \quad (30)$$

parce que le $\bar{\theta}$ pour \mathbf{x} est la matrice identité. En vertu de (30),

$$(\bar{x}_c, \bar{y}_c) = N \left[\sum_{i=1}^n \pi_i^{-1} \pi^{[x]_{\text{int}}} \right]^{-1} (\bar{x}_{p\pi}, \bar{y}_{p\pi})$$

$$= (\bar{x}_{p\pi}, \bar{y}_{p\pi}) + O_p(n^{-1}). \quad (31)$$

Donc,

$$\bar{y}_{p\pi} = \bar{y}_c + (\bar{x}_N - \bar{x}_c) \bar{\beta}_{c,1}$$

$$= \bar{y}_{p\pi} + (\bar{x}_N - \bar{x}_{p\pi}) \bar{\beta}_{c,1} + (\bar{y}_c - \bar{y}_{p\pi}) + (\bar{x}_{p\pi} - \bar{x}_c) \bar{\beta}_{c,1}$$

$$= \bar{y}_{p\pi} + O_p(n^{-1})$$

$$= \bar{y}_{\text{HT}} + (\bar{x}_N - \bar{x}_{\text{HT}}) \bar{\theta} + O_p(n^{-1}),$$

en vertu de (30), (31) et (29).

Remerciements

Cette étude a été financée aux termes de la coopération 43-3AEU-3-80088 entre la Iowa State University, le USDA National Agricultural Statistics Service et le U.S. Bureau of the Census, et aux termes de l'entente de coopération 68-3A75-14 entre le USDA

la fonction objective pour les MCO (PQ) en vue de produire des poids égaux ou supérieurs à 0 et inférieur à 0,065 accroît la somme moyenne des carrés de moins de 1 %. Voir le tableau 7. L'utilisation de la programmation quadratique pour imposer aux poids de régression SCP (SCP (PQL)) la borne zéro augmente très peu la somme moyenne des carrés, car seul un très petit nombre de poids sont modifiés.

Tableau 7

Moyenne de Monte Carlo de la somme des carrés des poids pour les échantillons ayant au moins un poids MCO négatif

Rég.	SCP	Rég.	SCP	Rég.	SCP	Rég.	SCP
MCO	PQ	SCP (PQL)	EMV	MCO	PQ	SCP (PQL)	EMV
1,208	1,217	1,226	1,227	1,342	1,342	1,242	

Le tableau 8 donne l'EQM de Monte Carlo pour les 562 échantillons ayant des poids par les moindres carrés ordinaires négatifs. La programmation quadratique est supérieure aux autres procédures dominant des poids non négatifs pour le percentile 0,01 et inférieur pour le percentile 0,99. Parmi les 562 échantillons, 497 avaient une moyenne d'échantillon supérieure à la moyenne de population. Si l'on échantillonnait une population finie, la borne sur les poids serait égale ou supérieure à N^{-1} et l'EQM de la programmation quadratique pour le percentile 0,99 serait réduite.

Tableau 8

EQM relative de Monte Carlo des estimateurs des percentiles pour les échantillons ayant au moins un poids MCO négatif

Percentile	MCO	PQ	SCP (PQL)	EMV	Rég.	MIO
0,01	287,52	291,11	350,58	461,80	344,06	72,50
0,05	76,04	70,88	75,80	88,71	38,84	36,05
0,10	44,80	40,74	39,31	38,84	12,56	3,45
0,25	20,24	19,14	14,72	9,91	3,35	0,75
0,50	5,03	4,53	3,65	2,26	0,90	0,95
0,75	5,02	4,53	3,66	4,24	14,56	14,56
0,90	23,77	23,69	20,04	18,80	20,49	37,94
0,95	51,54	46,04	30,79	28,28	32,54	205,85
0,99	206,33	90,08	39,40	57,54	43,49	235,71

La programmation quadratique est supérieure aux autres procédures calées pour le percentile 0,01 dans les échantillons avec poids par les MCO négatifs. La régression par la MIO et la régression pondérée SCP sont supérieures à l'EMV pour les percentiles 0,01 et 0,05. Il en est ainsi parce que l'EMV produit souvent le poids maximal le plus grand. Pour 3 026 des 30 000 échantillons, au moins un des poids de régression par la MIO est supérieur à 0,065. Pour

Donc, l'erreur quadratique moyenne relative de l'estimateur par les MCO du percentile 0,01 est de 283,27 %. Alors que cet estimateur du percentile 0,01 possède le biais le plus important, il donne l'erreur quadratique moyenne la plus faible parmi les procédures sans contrainte de borne. Les procédures PQ, MCO et logit donnent de meilleurs résultats pour les percentiles extrêmes, tandis que les autres sont meilleures pour les percentiles du milieu.

Tableau 5

Biais normalisé de Monte Carlo dans les estimateurs des percentiles

Percentile	MCO	PQ	SCP	EMV	Rég.	MIO	Logit
------------	-----	----	-----	-----	------	-----	-------

0,01	-7,75	-8,43	-2,88	-2,13	-4,70	-8,30	
0,05	-7,27	-7,95	-2,58	-1,82	-4,30	-7,85	
0,10	-6,66	-7,31	-2,27	-1,57	-3,91	-7,26	
0,25	-5,25	-5,82	-1,79	-1,25	-3,13	-5,89	
0,50	-3,21	-3,46	-1,37	-1,16	-2,18	-3,53	
0,75	-2,30	-2,07	-1,60	-2,21	-2,25	-1,78	
0,90	4,60	5,31	1,27	2,62	9,52	13,15	
0,95	12,75	13,33	6,01	6,41	26,65	30,03	
0,99	32,94	32,36	19,03	22,66			

Tableau 6

EQM relative de Monte Carlo des estimateurs des percentiles

Percentile	MCO	PQ	SCP	EMV	Rég.	MIO	Logit
0,01	283,27	282,50	30,23	311,58	296,37	282,76	
0,05	53,91	54,23	57,41	57,07	54,97	54,06	
0,10	25,50	25,97	26,40	25,79	25,26	25,80	
0,25	8,00	8,41	7,77	7,23	7,42	8,41	
0,50	1,99	2,07	1,88	1,71	1,83	2,12	
0,75	3,65	3,68	3,62	3,66	3,63	3,67	
0,90	14,50	14,60	14,25	14,57	14,36	14,56	
0,95	39,40	38,65	40,99	41,66	39,93	37,94	
0,99	200,17	196,24	235,71	216,22	205,85	194,33	

Dans 562 des 30 000 échantillons, au moins un des poids de régression par les MCO est négatif. Dans 17 échantillons, au moins un des poids de régression SCP originaux était négatif. L'utilisation de la programmation quadratique avec

$\alpha_i = \phi_{ii}^{11} = n^{-1}$. Les poids de l'estimateur par la MIQ et de l'EMV sont obtenus en minimisant les fonctions objectives (5) et (7), respectivement, sous la contrainte (4). Les poids pour l'estimateur par la régression pondérée SCP sont donnés par (22). Les poids sont représentés graphiquement en fonction des valeurs x d'échantillon à la figure 1. Cinq des poids de régression simple sont négatifs, à cause de l'écart important entre les moyennes d'échantillon et de population. Pour l'estimateur par la régression pondérée SCP, l'EMV et la MIQ, tous les poids sont non négatifs. La figure 1 montre que les comportements des poids obtenus par la MIQ et par la régression pondérée SCP sont semblables et que l'EMV produit un poids extrêmement grand dans cet échantillon.

Le tableau 1 contient certains poids pour les valeurs de x les plus petites, les valeurs de x proches de la moyenne d'échantillon, les valeurs de x les plus grandes. Pour les valeurs de x les plus éloignées de la moyenne de population, l'EMV donne les poids les plus grands. Pour les valeurs de x les plus proches de la moyenne de population, les poids de l'EMV sont proches de n^{-1} , tandis que les autres poids sont grands.

Étude par simulation

Afin de comparer les diverses méthodes de construction des poids de régression, nous avons réalisé un étude par simulation. En tout, nous avons sélectionné 30 000 échantillons aléatoires simples de taille 32 à partir d'une loi χ^2 à

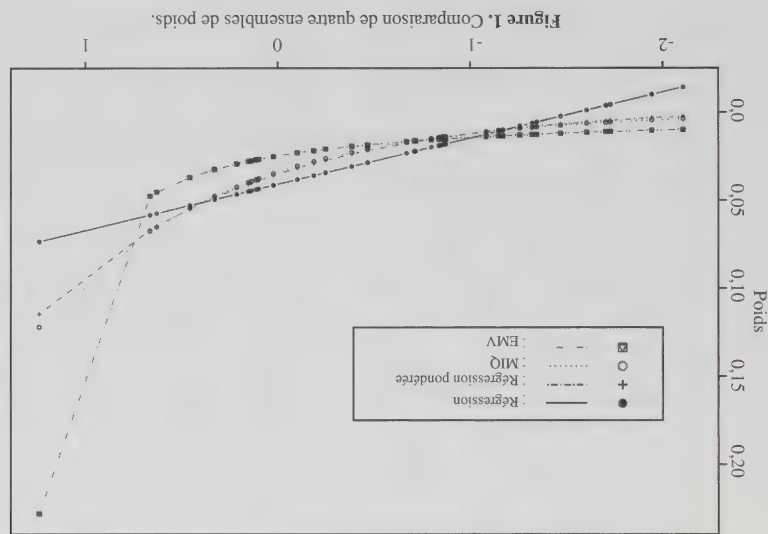


Figure 1. Comparaison de quatre ensembles de poids.

Tableau 1				
Certains poids de régression pour l'exemple illustré				
Poids multipliés par $n = 40$				
x	Rég.	Rég.	MIQ	EMV
		pond.		
-2,103	-0,56	0,12	0,16	0,40
-1,941	-0,40	0,12	0,20	0,40
-1,727	-0,16	0,20	0,24	0,44
-0,710	0,88	0,68	0,68	0,68
-0,670	0,96	0,72	0,68	0,68
-0,468	1,16	0,88	0,84	0,76
-0,103	1,52	1,28	1,24	0,92
0,021	1,68	1,44	1,40	1,00
0,097	1,76	1,56	1,52	1,08
0,628	2,32	2,60	2,60	1,84
0,662	2,36	2,68	2,72	1,92
1,237	2,96	4,60	4,88	9,12

1. Régression par les moindres carrés ordinaires (MCO)
2. Programmation quadratique avec bornes supérieure et inférieure (PQ)
3. Régression pondérée avec poids SCP (Rég. SCP)
4. Fonction objective pour le maximum de vraisemblance (EMV)
5. Fonction objective pour la MIQ (Rég. MIQ))
6. Procédure logit avec bornes supérieure et inférieure (Logit)

où

$$\hat{\mathbf{G}}_{\mathbf{x}\mathbf{x},(t)} = (\mathbf{x}_{\text{HT}} - \mathbf{x}_N - \mathbf{d}_{x_i})' \sum_{i=1}^{x-1} \hat{\pi}_{i|\mathbf{x}_{\text{HT}}}^{-1} (\mathbf{x}_{\text{HT}} - \mathbf{x}_N - \mathbf{d}_{x_i});$$

Poisons que l'estimateur (12) construit avec les $\hat{\pi}_{i|\mathbf{x}_{\text{HT}}}$ de

$$(21) \quad \underline{y}^{\text{preg}} = N^{-1} \sum_{i=1}^I \hat{\pi}_{i|\mathbf{x}_{\text{HT}}}^{-1} y_i.$$

La probabilité de sélection conditionnelle approximative pour un échantillonnage aléatoire simple et une seule variable auxiliaire est donnée par

$$\hat{\pi}_{i|\mathbf{x}_i} = \frac{N}{n} \left[\frac{\hat{\varpi}_{\mathbf{x}}}{\hat{\varpi}_{\mathbf{x},(t)}} \right]$$

$$\left\{ \exp \left[\frac{1}{2} \frac{\hat{\varpi}_{\mathbf{x}}^2}{(\mathbf{x}_i^n - \mathbf{x}_N)^2} - \frac{\hat{\varpi}_{\mathbf{x},(t)}^2}{(\mathbf{x}_i^n - \mathbf{x}_N - \mathbf{d}_{x_i})^2} \right] \right\},$$

où

$$d_{x_i} = [n(N-1)^{-1}(N-n)(x_i - x_N)],$$

$$\hat{\varpi}_{\mathbf{x},(t)}^2 = \frac{(N-n)(n-1)}{(N-2)} \left[s_{\mathbf{x}}^2 - \frac{N(x_i - x_N)^2}{n-1} \hat{\varpi}_{\mathbf{x}}^2 \right] \approx \frac{n}{n-1} \hat{\varpi}_{\mathbf{x}}^2,$$

et

$$s_{\mathbf{x}}^2 = (n-1)^{-1} \sum_{i=1}^I (x_i - \bar{x})^2.$$

Dans ce cas, $d_{x_i} = \bar{x}_N^{(t)} - x_N$ et $M_{\mathbf{x}\mathbf{y}} = \text{Cov}(\bar{\mathbf{x}}_{\text{HT}}, \bar{y}_{\text{HT}})$.

L'estimateur SCP (21) obtenu avec les probabilités de sélection conditionnelles approximatives n'est pas calé; autrement dit l'estimateur (21) de la moyenne du vecteur de variables auxiliaires n'est pas le vecteur de moyennes de population. Il est assez facile de normaliser les probabilités de sorte que leur somme soit égale à l'unité ou à la fraction de strate en cas d'échantillonnage stratifié. Pour construire un estimateur calé pour le cas général, nous proposons de calculer l'estimateur par la régression avec $[\sum_{j=1}^I \hat{\pi}_{j|\mathbf{x}_{\text{HT}}}^{-1}]^{-1}$ comme poids initiaux. L'estimateur proposé est

$$\underline{y}^{\text{preg}} = \underline{y}_c + (\mathbf{x}_N - \mathbf{x}_c) \hat{\beta}_{c,1}$$

où

$$(\underline{y}_c, \mathbf{x}_c) = \sum_{i=1}^I \alpha_i (y_i, \mathbf{x}_i),$$

$$(\hat{\beta}_{c,0}, \hat{\beta}_{c,1})' = \left[\sum_{i=1}^I \alpha_i' \mathbf{z}_i' \mathbf{z}_i \right]^{-1} \left[\sum_{i=1}^I \alpha_i' \mathbf{z}_i' y_i \right],$$

$$\mathbf{z}_i = (1, \mathbf{x}_i - \mathbf{x}_c),$$

$$w_i = \alpha_i \left[\sum_{j=1}^f \alpha_j (\mathbf{x}_j - \mathbf{x}_c) (\mathbf{x}_j - \mathbf{x}_c)' \right]^{-1} \alpha_i (\mathbf{x}_i - \mathbf{x}_c)',$$

et $\hat{\pi}_{i|\mathbf{x}_{\text{HT}}}$ est la probabilité de sélection conditionnelle approximative donnée par (20). Nous supposons que le vecteur de variables auxiliaires contient la valeur n si bien que l'estimateur ne varie pas en fonction de la localisation.

L'estimateur (21) est approximativement égal à un estimateur par la régression et l'estimateur (22) est, lui aussi approximativement égal au même estimateur par la régression.

Théorème : Soit une série de populations et d'échantillons, $\{F_N, A_N\}$, satisfaisant

$$(23) \quad (\underline{y}_{\text{HT}}, \mathbf{x}_{\text{HT}}) - (\underline{y}_N, \mathbf{x}_N) = O_p(n^{-1/2}).$$

Supposons que les séries de matrices des covariances estimes, $\sum_{\mathbf{x}\mathbf{x}}^{\mathbf{x}\mathbf{x}}$ et $\sum_{\mathbf{x}\mathbf{x},(t)}^{\mathbf{x}\mathbf{x}}$, satisfont

$$(24) \quad - \left[\mathbf{D}^{-1/2} \sum_{\mathbf{x}\mathbf{x}}^{\mathbf{x}\mathbf{x}} \mathbf{D}^{-1/2} \right]^{-1} = O_p(n^{-1}),$$

où \mathbf{D} représente une matrice diagonale ayant sur sa diagonale les éléments de $\sum_{\mathbf{x}\mathbf{x}}^{\mathbf{x}\mathbf{x}}$. Soit \mathbf{d}_{x_i}

une fonction de l'échantillon satisfaisant (19) et supposons que (18) est vérifiée. Supposons que la série d'estimateurs de la variance d'Horvitz-Thompson satisfait

$$(25) \quad \text{Var} \left\{ n \left[\text{Vech} \left(\sum_{\mathbf{z}\mathbf{z}, \text{HT}}^{\mathbf{z}\mathbf{z}} - \sum_{\mathbf{z}\mathbf{z}}^{\mathbf{z}\mathbf{z}} \right) \right] \right\} = O(n^{-1}),$$

où $\mathbf{z}_i = (\mathbf{x}_i', y_i')$ et $\sum_{\mathbf{z}\mathbf{z}}^{\mathbf{z}\mathbf{z}}$ est définie positive. Supposons que $E[\hat{\pi}_{i|\mathbf{x}_{\text{HT}}}^{-2}]$ est bornée, où $\hat{\pi}_{i|\mathbf{x}_{\text{HT}}}$ est défini dans (20). Alors, l'estimateur SCP $\underline{y}^{\text{preg}}$ de (21) satisfait

$$\underline{y}^{\text{preg}} = \underline{y}_{\text{HT}} + (\mathbf{x}_N - \mathbf{x}_{\text{HT}}) \hat{\theta} + O_p(n^{-1})$$

$$\hat{\theta} = \sum_{i=1}^N \hat{\mathbf{M}}_{\mathbf{x}\mathbf{x}}^{\mathbf{x}\mathbf{x}} \text{ et } \theta_N = \sum_{i=1}^N \mathbf{M}_{\mathbf{x}\mathbf{x}}^{\mathbf{x}\mathbf{x}}.$$

Si $\text{Var} \{ \sum_{i=1}^I \pi_i^{-1} \pi_i' \} > 0$, supposons que \mathbf{x}_i contient la valeur un commun élément. Supposons que $\mathbf{M}_{\mathbf{x}\mathbf{y}}^{\mathbf{x}\mathbf{y}} = \sum_{\mathbf{x}\mathbf{y}}^{\mathbf{x}\mathbf{y}}$ satisfait

$$\underline{y}^{\text{preg}} = \underline{y}_{\text{HT}} + (\mathbf{x}_N - \mathbf{x}_{\text{HT}}) \hat{\theta} + O_p(n^{-1}).$$

Pour la preuve, voir l'annexe. Afin d'illustrer la nature des divers types de poids de régression, nous avons tiré un échantillon simple de taille 40 à partir d'une population normale de moyenne nulle et de variance égale à un. La moyenne d'échantillon est -0,614 et la moyenne de population est nulle. Les poids de l'estimateur par la régression sont donnés par (2) avec

production de poids de régression qui sont non négatifs avec probabilité élevée. Supposons que le vecteur des moyennes de population des variables auxiliaires, \mathbf{x}_N , soit connu. Considérons l'estimateur d'Horvitz-Thompson de \mathbf{x}_N donné par

$$\bar{\mathbf{x}}_{\text{HT}} = \frac{1}{N} \sum_{i=1}^N \frac{\mathbf{x}_i}{\pi_i}, \quad (11)$$

où $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ et π_i est la probabilité de sélection inconditionnelle. Tillé (1998) propose l'estimateur simple

$$\bar{y}^{\text{pr}} = \frac{1}{N} \sum_{i=1}^N \frac{y_i}{\pi_i}, \quad (12)$$

où $\pi_{i|\mathbf{x}_{\text{HT}}}$ est la probabilité de sélection conditionnelle du i^{e} élément, sachant \mathbf{x}_{HT} . Pour construire l'estimateur SCP de \bar{y}_N , il faut connaître la probabilité de sélection conditionnelle $\pi_{i|\mathbf{x}_{\text{HT}}}$. Si \mathbf{x}_{HT} prend la valeur \mathbf{t} , nous avons

$$\pi_{i|\mathbf{x}_{\text{HT}}} = \pi_i \frac{P\{\mathbf{x}_{\text{HT}} = \mathbf{t} | i \in A\}}{P\{\mathbf{x}_{\text{HT}} = \mathbf{t}\}}, \quad (13)$$

où A est l'ensemble d'indices pour les éléments de l'échantillon.

Afin de calculer les probabilités de sélection conditionnelles, il faut connaître la loi de probabilité de \mathbf{x}_{HT} inconditionnelle et conditionnelle à la présence de chaque unité dans l'échantillon. À part certains cas particuliers, cette loi de probabilité est fort complexe. Par conséquent, nous considérons l'approximation de la probabilité de sélection conditionnelle.

Sous l'hypothèse que \mathbf{x}_{HT} suit une loi approximativement normale inconditionnellement et conditionnellement à la présence de chaque unité dans l'échantillon, la probabilité de sélection inconditionnelle (13) peut être

approximée par

$$\pi_{i|\mathbf{x}_{\text{HT}}} = \pi_i \left| \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \right|^{1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \mathbf{x}_{\text{HT}})' \Sigma_{\mathbf{x}\mathbf{x}}^{-1} (\mathbf{x}_i - \mathbf{x}_{\text{HT}}) \right\}, \quad (14)$$

où $\Sigma_{\mathbf{x}\mathbf{x}} = \text{Var}\{\mathbf{x}_{\text{HT}} | F\}$, $\Sigma_{\mathbf{x}\mathbf{x},(i)} = \text{Var}\{\mathbf{x}_{\text{HT}} | F, i \in A\}$,

$$\mathbf{G}_{\mathbf{x}\mathbf{x}} = (\mathbf{x}_{\text{HT}} - \mathbf{x}_N)' \Sigma_{\mathbf{x}\mathbf{x}}^{-1} (\mathbf{x}_{\text{HT}} - \mathbf{x}_N),$$

$$\mathbf{G}_{\mathbf{x}\mathbf{x},(i)} = (\mathbf{x}_{\text{HT}} - \mathbf{x}_{N,(i)})' \Sigma_{\mathbf{x}\mathbf{x},(i)}^{-1} (\mathbf{x}_{\text{HT}} - \mathbf{x}_{N,(i)}),$$

$$\mathbf{x}_{N,(i)} = E\{\mathbf{x}_{\text{HT}} | F, i \in A\} =$$

$$(N\pi_i)^{-1} \mathbf{x}_i + N^{-1} \sum_{j=1, j \neq i}^f (\pi_i \pi_j)^{-1} \pi_j \mathbf{x}_j,$$

A est l'ensemble d'indices qui apparaissent dans l'échantillon et $F = \{y_1, \dots, y_N\}$ est la population finie. Tillé (1998) donne une expression pour $\Sigma_{\mathbf{x}\mathbf{x},(i)}$ pour le cas général.

Supposons que les matrices des covariances de plan de sondage $\Sigma_{\mathbf{x}\mathbf{x}}$ et $\Sigma_{\mathbf{x}\mathbf{x},(i)}$ sont définies positives et que le vecteur de variables auxiliaires suit une loi normale. Tillé (1999) montre que l'estimateur SCP défini en (12) avec les probabilités de sélection conditionnelles approximatives données par (14) satisfait

$$\bar{y}^{\text{pr}} = \bar{y}_{\text{HT}} - \mathbf{x}_{\text{HT}}' \beta_N + O_p(n^{-1}), \quad (15)$$

$$= \bar{y}^{\text{reg}} + O_p(n^{-1}), \quad (16)$$

où

$$\beta_N = \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \Sigma_{\mathbf{x}\mathbf{y}},$$

$$\bar{y}^{\text{reg}} = \bar{y}_{\text{HT}} + (\mathbf{x}_N - \mathbf{x}_{\text{HT}})' \beta,$$

$$\beta = \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \Sigma_{\mathbf{x}\mathbf{y}} = (\mathbf{X}' \Phi^{-1} \mathbf{X})^{-1} \mathbf{X}' \Phi^{-1} \mathbf{y},$$

$\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_N)'$, $\mathbf{y} = (y_1, \dots, y_N)'$, le i^{e} élément de Φ^{-1} est $N^{-2}(\pi_i \pi_j)^{-1}(\pi_j - \pi_i \pi_j)$, $\Sigma_{\mathbf{x}\mathbf{x}}$ est la variance par rapport au plan de sondage de \mathbf{x}_{HT} , $\Sigma_{\mathbf{x}\mathbf{y}}$ est la covariance par rapport au plan de sondage de \mathbf{x}_{HT} et \bar{y}_{HT} , $\Sigma_{\mathbf{x}\mathbf{x}}$ est l'estimateur d'Horvitz-Thompson de la variance de \mathbf{x}_{HT} , et $\Sigma_{\mathbf{x}\mathbf{y}}$ est l'estimateur d'Horvitz-Thompson de la covariance de \mathbf{x}_{HT} et \bar{y}_{HT} . Dans le cas d'un plan de sondage complexe, un certain nombre de quantités figurant dans (14) sont difficiles à calculer. Cependant, les approximations des estimateurs donnant les mêmes propriétés en grand échantillon sont assez faciles à calculer. Nous remplaçons $\Sigma_{\mathbf{x}\mathbf{x}}$ et $\Sigma_{\mathbf{x}\mathbf{y},(i)}$ par les estimateurs, nous remplaçons $\mathbf{x}_{N,(i)}$ par \mathbf{x}_N et \mathbf{d}_{x_i} nous définissons

$$\hat{\mathbf{M}}_{\mathbf{x}\mathbf{y}} = \sum_{i \in A} (N\pi_i)^{-1} \mathbf{d}_{x_i} y_i, \quad (17)$$

et nous supposons que

$$\text{Var}\{n(\hat{\mathbf{M}}_{\mathbf{x}\mathbf{y}} - \mathbf{M}_{\mathbf{x}\mathbf{y}})\} = O(n^{-1}), \quad (18)$$

$$\mathbf{d}_{x_i} = O_p(n^{-1}), \quad (19)$$

où \mathbf{d}_{x_i} est une fonction de l'échantillon et $\mathbf{M}_{\mathbf{x}\mathbf{y}}$ est une quantité de population. Souvent, $\mathbf{M}_{\mathbf{x}\mathbf{y}}$ est la matrice des covariances de population $\Sigma_{\mathbf{x}\mathbf{y}}$, mais cette égalité n'est pas nécessaire pour que l'estimateur soit bien défini. Dans de nombreux cas, on peut calculer \mathbf{d}_{x_i} sous forme d'un multiple de l'écart jackknife. En outre, dans de nombreuses situations, une valeur adéquate de l'estimateur, $\Sigma_{\mathbf{x}\mathbf{x},(i)}$, de $\Sigma_{\mathbf{x}\mathbf{x},(i)}$ est $n^{-1}(n-1)\Sigma_{\mathbf{x}\mathbf{x}}$. Nous écrivons notre généralisation de (14) comme suit

$$\pi_{i|\mathbf{x}_{\text{HT}}} = \pi_i \left| \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \right|^{1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \mathbf{x}_{\text{HT}})' \Sigma_{\mathbf{x}\mathbf{x}}^{-1} (\mathbf{x}_i - \mathbf{x}_{\text{HT}}) \right\}, \quad (20)$$

sont données pour la différence entre l'estimateur final d'une partie du vecteur de variables auxiliaires et les éléments correspondants du vecteur de population.

Dans le présent article, nous considérons divers types de poids de régression, y compris une procédure fondée sur les probabilités de sélection conditionnelles de Tillé (1998). Nous utilisons les probabilités de sélection conditionnelles pour calculer des poids de régression qui sont positifs pour la plupart des échantillons. Nous comparons ces poids à ceux obtenus par la MIQ, la programmation quadratique, une procédure logit et l'estimation du maximum de vraisemblance.

2. Maximum de vraisemblance et MIQ

Considérons un tableau à double entrée contenant r lignes et c colonnes. La cellule de population U_{ij} contient

N_{ij} éléments; $i = 1, \dots, r$, $j = 1, \dots, c$. Supposons que les dénombrements de marge $N_{i\cdot}$, $N_{\cdot j}$ soient connus. Les caractéristiques de la population d'intérêt sont les N_{ij} ou, de façon équivalente, les $p_{ij} = N_{ij}^{-1} N_{\cdot j}$. Pour un échantillon aléatoire simple sans remise de taille n , Deming et Stephan (1940) ont proposé une méthode d'ajustement proportionnel itératif appelée méthode itérative du quotient

(MIQ) pour obtenir la solution pour les fréquences de cellule. Voir aussi Stephan (1942). Si nous supposons que l'échantillon est un échantillon aléatoire issu d'une loi multinomiale définie par les valeurs de population dans un tableau à double entrée, nous pouvons construire un estimateur par la méthode du maximum de vraisemblance.

Deville et Särndal (1992) ont défini une classe d'estimateurs par calage, \bar{y}_{cal} , de la moyenne de population de y de la forme

$$\bar{y}_{cal} = \sum_{i=1}^n w_i y_i, \quad (3)$$

où les w_i minimisent la fonction objective $\sum_{i=1}^n G(w_i, \alpha_i)$ sous les contraintes

$$\sum_{i=1}^n w_i = 1 \quad \text{et} \quad \sum_{i=1}^n w_i x_i = \bar{x}_N, \quad (4)$$

et $G(w_i, \alpha_i)$ est une mesure de la distance entre le poids initial α_i et le poids final w_i . Les estimateurs par la MIQ et par le maximum de vraisemblance de la fraction de population de la cellule, p_{ij} , appartiennent à la classe des estimateurs par calage.

Pour un échantillon aléatoire simple, nous pouvons obtenir les poids par la MIQ pour la fraction de population de la cellule en minimisant

$$\sum_{i=1}^n w_i \log \left(\frac{n-1}{w_i} \right) - (n-1) \log \left(\frac{n-1}{\bar{w}_k} \right) \quad (5)$$

sous les contraintes (4) avec

$$\bar{w}_k = \frac{1}{n} \sum_{i=1}^n w_i \log \left(\frac{n-1}{w_i} \right) - (n-1) \log \left(\frac{n-1}{\bar{w}_k} \right) \quad (6)$$

ou les contraintes (4) avec $x_i = \bar{x}_N$.

Deville et Särndal (1992) montrent que les estimateurs par la MIQ et par le maximum de vraisemblance équivalent approximativement à un estimateur par la régression de la forme (1) et, par conséquent, ont la même loi limite que l'estimateur par la régression. Les poids pour les estimateurs par la MIQ et par le maximum de vraisemblance sont non négatifs si les solutions pour les poids existent.

3. Régression pondérée en utilisant des probabilités conditionnelles

Tillé (1998) propose d'utiliser des probabilités de sélection conditionnelle approximatives, sachant les estimateurs d'Horvitz-Thompson des variables auxiliaires, pour calculer un estimateur de la moyenne de population de la variable étudiée. Son approximation peut être étendue à la

diverses procédures en vue de construire des vecteurs de poids de régression non négatifs sont considérées. Un vecteur de poids de régressions dans lequel les poids initiaux sont les inverses des probabilités de sélection conditionnelles approximatives est présenté. Une étude par simulation permet de comparer les poids obtenus par la régression pondérée, la programmation quadratique, la méthode itérative du quotient, une procédure logit et la méthode du maximum de vraisemblance.

Mois clés : Méthode itérative du quotient; maximum de vraisemblance; programmation quadratique; estimateur simple conditionnellement pondéré.

1. Introduction

Dans le domaine du sondage, on dispose souvent d'information au sujet de la population à l'étape de l'analyse. L'estimation par la régression est l'une des méthodes choisies pour utiliser cette information. La construction d'un estimateur par la régression d'une moyenne ou d'un total de population peut se faire de plusieurs façons. L'un des estimateurs de la moyenne par la régression est

$$(I) \quad \tilde{g}({}^u \underline{x} - {}^N \underline{x}) + {}^u \underline{\kappa} = {}^I \underline{\kappa} \quad {}^I \mathcal{M} \sum_y {}^I = {}^I \text{reg} \underline{\kappa}$$

no

$$(7) \quad {}^1_{l-} \phi^t \mathbf{x} \left({}^f_{l-} \mathbf{x} {}^f_{l-} \phi^t \mathbf{x} \sum_u^{l=f} \right) ({}^u \mathbf{x} - {}^N \mathbf{x}) + {}^t \mathbf{x} = {}^l \mathbf{x}$$

$$({}^I \mathbf{x} \circ {}^I \mathbf{A}) \circ \sum_{u=1}^I := ({}^I \mathbf{x} \circ {}^I \mathbf{A}) \circ {}^I \mathbf{u} \sum_{u=1}^I \left({}^I \mathbf{u} \sum_{u=1}^I \right) = ({}^u \mathbf{x} \circ {}^u \mathbf{A})$$

$${}^t\mathcal{K} \, {}^t\mathbf{I}^{\text{II}}\Phi \, {}^t\mathbf{X} \sum_u^{I=1} \left({}^t\mathbf{X} \, {}^t\mathbf{I}^{\text{II}}\Phi \, {}^t\mathbf{X} \sum_u^{I=1} \right) = \mathfrak{g}$$

$$e_{1-}^{\dagger} u \left(\begin{matrix} f \\ 1- \\ u \end{matrix} \sum_{u=1}^f \right) = {}^f \chi$$

est la moyenne de population de \mathbf{x} . Un choix possible pour singulière, les π_i sont les probabilités de sélection et \mathbf{x}^N $\Phi = \text{diag}(\phi_1^{uu}, \dots, \phi_m^{uu})$ est une matrice diagonale non

Comme dans une table de fréquences.

Une autre modification des poids de régression consiste à assembler certaines contraintes appliquées pour construire l'estimateur. Husain (1969) envisage de modifier les poids d'un échantillon aléatoire simple issu d'une loi normale. Il calcule les poids qui minimisent l'erreur quadratique moyenne (EQM) de l'estimateur résultant. Bardsley et Chambers (1984) considèrent un estimateur fondé sur une fonction objective et sur la division de la variable auxiliaire en deux composantes. Ils étudient le comportement de l'estimateur dans la perspective d'un modèle. Rao et Singh (1997) étudient un estimateur dans lequel des tolérances

utilisés pour estimer un total de population finit dans le cas d'une enquête générale, il semble raisonnable de poser qu'aucun poids individuel ne soit inférieur à 1. En outre, il semble raisonnable, du point de vue de la robustesse, d'éviter les poids dont la valeur est très grande.

Il existe plusieurs moyens de construire des poids de régression dont la fourchette de valeur est réduite. Hwang et Fuller (1978) définissent un procédé pour modifier les w_i de sorte qu'il n'y ait aucun poids négatif et aucun poids de valeur élevée. Husain (1969) propose de recourir à la programmation quadratique pour imposer des bornes aux poids. La programmation quadratique et plusieurs autres procédés reposent sur le fait qu'on peut définir les poids comme des valeurs qui optimisent une certaine fonction. Deville et Särndal (1992) envisagent sept fonctions objectives susceptibles d'être utilisées pour construire les poids. Ils proposent des fonctions objectives utilisables pour produire des poids qui tombent dans une fourchette donnée. Deville, Särndal et Sautory (1993) présentent le programme CALMAR, rédigé sous forme de macro SAS, qui peut être utilisé pour calculer les poids correspondant à quatre fonctions objectives distinctes, quand l'information auxiliaire dans l'enquête correspond à des dénombrements de marges

- Nandram et Choi : Modèles hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines
- Squires, S. (2001). National plan urges to combat obesity: Weight-related illnesses kill 300,000 Americans annually. Surgeon General says. *The Washington Post*, 14 décembre, 2001.
- Slasny, E.A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the National Crime Survey. *Journal of the American Statistical Association*, 86, 296-303.
- Who Consultation on Obesity (2000). Obesity: Preventing and managing the global epidemic. *WHO Technical Report Series* 894. Geneva, Switzerland: World Health Organization.
- Wright, C.M., Parker, L., Lamont, D. et Craft, A.W. (2001). Implications of childhood obesity for adult health: Findings from thousand families cohort study. *British Medical Journal*, 323, 1280-1284.
- Ruben, H. (1966). Some new results on the distribution of the sample correlation coefficient. *Journal of the Royal Statistical society, Series B*, 28, 513-525.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.
- Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.
- Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J. A. et Rubin, D.B. (1996). The NHANES III multiple imputation project. *Survey research methods, Proceedings of the American Statistical Association*, 28-37.

Tableau 3
Comparaison des intervalles de sélection (régression avec spline et régression sans spline) fondée sur la moyenne des bornes des intervalles de confiance à 95 % sur l'ensemble des comités pour la moyenne en population finie de l'IMC selon l'âge, la race et le sexe

R-S	âge	Note : R-S = race-sexe, FN = femme noire, MN = homme noir, AF = autre femme, et AM = autre homme, non					
		15 à 19 ans	10 à 14 ans	5 à 9 ans	2 à 4 ans	10 à 14 ans	15 à 19 ans
FN	Pas de spline	(16,26, 16,92)	(16,44, 17,10)	(16,62, 18,41)	(17,62, 18,41)	(16,26, 16,92)	(15,65, 16,31)
	Spline	(15,65, 16,31)	(16,44, 17,10)	(16,26, 16,92)	(17,62, 18,41)	(16,26, 16,92)	(15,65, 16,31)
MN	Pas de spline	(16,10, 16,76)	(16,26, 16,92)	(16,26, 16,92)	(16,26, 16,92)	(16,10, 16,76)	(15,68, 16,32)
	Spline	(15,68, 16,32)	(16,26, 16,92)	(16,26, 16,92)	(16,26, 16,92)	(15,68, 16,32)	(15,68, 16,32)
AF	Pas de spline	(16,39, 17,00)	(16,56, 17,17)	(16,56, 17,17)	(16,56, 17,17)	(16,39, 17,00)	(21,16, 25,39)
	Spline	(16,01, 16,60)	(16,56, 17,17)	(16,56, 17,17)	(16,56, 17,17)	(16,01, 16,60)	(21,16, 25,39)
AM	Pas de spline	(16,53, 17,14)	(16,67, 17,29)	(16,67, 17,29)	(16,67, 17,29)	(16,53, 17,14)	(20,83, 24,98)
	Spline	(16,16, 16,74)	(16,67, 17,29)	(16,67, 17,29)	(16,67, 17,29)	(16,16, 16,74)	(20,83, 24,98)

toutes les autres hypothèses demeurent par ailleurs inchangées. Ce modèle ne présente aucune amélioration appréciable par rapport au modèle spécifié par (9), que nous retenons sans autre perfectionnement.

Au tableau 3, nous comparons la MPF pour les modèles de sélection (régression sans spline et régression avec spline). De nouveau, nous calculons la moyenne des bornes des intervalles de confiance à 95 % sur l'ensemble des comités. Les intervalles se chevauchent, ce qui porte à croire qu'il existe une similitude entre les modèles avec et sans spline. Cependant, nous notons certaines exceptions. L'écart le plus important entre les intervalles a lieu pour les jeunes de 15 à 19 ans. En général, le modèle spline donne une plus grande précision. Par exemple, pour le groupe des 10 à 19 ans, les intervalles pour le modèle spline sont contenus dans ceux obtenus pour le modèle sans les spline.

6. Conclusions

Pour analyser les données sur l'IMC provenant de la NHANES III selon l'âge, la race et le sexe dans chaque comité, a) nous avons étendu le modèle de régression logistique-normale à deux modèles de sélection hiérarchique bayésien et b) construit un modèle de mélange de schémas d'observation et deux modèles à non-réponse ignorable pour évaluer la sensibilité à l'inférence. Une mesure de déviance montre que, des quatre modèles, le modèle de sélection est le meilleur et une analyse de vérification croisée montre que l'ajustement des modèles est à peu près équivalent.

Une autre contribution de l'étude est le dépistage d'une déficience commune au modèle de sélection, au modèle de mélange de schémas d'observation et aux deux modèles à non-réponse ignorable. D'après les données observées, nous avons constaté qu'il existe une relation dynamique entre l'IMC et l'âge. Par conséquent, nous avons étendu plus loin le modèle de sélection afin d'inclure trois splines linéaires.

Annexe A

Le modèle de mélange de schémas d'observation

Pour la partie I du modèle de mélange de schémas d'observation, la réponse dépend de l'âge, de la race et du sexe, ainsi que de l'interaction de la race et du sexe par la

voie de la régression logistique.

$$r_{ij} | \beta_i \sim \text{Bernoulli} \left\{ \frac{e^{\beta_{0i} + \beta_{1i}x_{ij} + \beta_{2i}x_{ij}^2 + \beta_{3i}x_{ij}^3 + \beta_{4i}x_{ij}^4 + \beta_{5i}x_{ij}^5 + \beta_{6i}x_{ij}^6 + \beta_{7i}x_{ij}^7 + \beta_{8i}x_{ij}^8 + \beta_{9i}x_{ij}^9 + \beta_{10i}x_{ij}^{10} + \beta_{11i}x_{ij}^{11} + \beta_{12i}x_{ij}^{12} + \beta_{13i}x_{ij}^{13} + \beta_{14i}x_{ij}^{14} + \beta_{15i}x_{ij}^{15} + \beta_{16i}x_{ij}^{16} + \beta_{17i}x_{ij}^{17} + \beta_{18i}x_{ij}^{18} + \beta_{19i}x_{ij}^{19} + \beta_{20i}x_{ij}^{20} + \beta_{21i}x_{ij}^{21} + \beta_{22i}x_{ij}^{22} + \beta_{23i}x_{ij}^{23} + \beta_{24i}x_{ij}^{24} + \beta_{25i}x_{ij}^{25} + \beta_{26i}x_{ij}^{26} + \beta_{27i}x_{ij}^{27} + \beta_{28i}x_{ij}^{28} + \beta_{29i}x_{ij}^{29} + \beta_{30i}x_{ij}^{30} + \beta_{31i}x_{ij}^{31} + \beta_{32i}x_{ij}^{32} + \beta_{33i}x_{ij}^{33} + \beta_{34i}x_{ij}^{34} + \beta_{35i}x_{ij}^{35} + \beta_{36i}x_{ij}^{36} + \beta_{37i}x_{ij}^{37} + \beta_{38i}x_{ij}^{38} + \beta_{39i}x_{ij}^{39} + \beta_{40i}x_{ij}^{40} + \beta_{41i}x_{ij}^{41} + \beta_{42i}x_{ij}^{42} + \beta_{43i}x_{ij}^{43} + \beta_{44i}x_{ij}^{44} + \beta_{45i}x_{ij}^{45} + \beta_{46i}x_{ij}^{46} + \beta_{47i}x_{ij}^{47} + \beta_{48i}x_{ij}^{48} + \beta_{49i}x_{ij}^{49} + \beta_{50i}x_{ij}^{50} + \beta_{51i}x_{ij}^{51} + \beta_{52i}x_{ij}^{52} + \beta_{53i}x_{ij}^{53} + \beta_{54i}x_{ij}^{54} + \beta_{55i}x_{ij}^{55} + \beta_{56i}x_{ij}^{56} + \beta_{57i}x_{ij}^{57} + \beta_{58i}x_{ij}^{58} + \beta_{59i}x_{ij}^{59} + \beta_{60i}x_{ij}^{60} + \beta_{61i}x_{ij}^{61} + \beta_{62i}x_{ij}^{62} + \beta_{63i}x_{ij}^{63} + \beta_{64i}x_{ij}^{64} + \beta_{65i}x_{ij}^{65} + \beta_{66i}x_{ij}^{66} + \beta_{67i}x_{ij}^{67} + \beta_{68i}x_{ij}^{68} + \beta_{69i}x_{ij}^{69} + \beta_{70i}x_{ij}^{70} + \beta_{71i}x_{ij}^{71} + \beta_{72i}x_{ij}^{72} + \beta_{73i}x_{ij}^{73} + \beta_{74i}x_{ij}^{74} + \beta_{75i}x_{ij}^{75} + \beta_{76i}x_{ij}^{76} + \beta_{77i}x_{ij}^{77} + \beta_{78i}x_{ij}^{78} + \beta_{79i}x_{ij}^{79} + \beta_{80i}x_{ij}^{80} + \beta_{81i}x_{ij}^{81} + \beta_{82i}x_{ij}^{82} + \beta_{83i}x_{ij}^{83} + \beta_{84i}x_{ij}^{84} + \beta_{85i}x_{ij}^{85} + \beta_{86i}x_{ij}^{86} + \beta_{87i}x_{ij}^{87} + \beta_{88i}x_{ij}^{88} + \beta_{89i}x_{ij}^{89} + \beta_{90i}x_{ij}^{90} + \beta_{91i}x_{ij}^{91} + \beta_{92i}x_{ij}^{92} + \beta_{93i}x_{ij}^{93} + \beta_{94i}x_{ij}^{94} + \beta_{95i}x_{ij}^{95} + \beta_{96i}x_{ij}^{96} + \beta_{97i}x_{ij}^{97} + \beta_{98i}x_{ij}^{98} + \beta_{99i}x_{ij}^{99} + \beta_{100i}x_{ij}^{100} + \beta_{101i}x_{ij}^{101} + \beta_{102i}x_{ij}^{102} + \beta_{103i}x_{ij}^{103} + \beta_{104i}x_{ij}^{104} + \beta_{105i}x_{ij}^{105} + \beta_{106i}x_{ij}^{106} + \beta_{107i}x_{ij}^{107} + \beta_{108i}x_{ij}^{108} + \beta_{109i}x_{ij}^{109} + \beta_{110i}x_{ij}^{110} + \beta_{111i}x_{ij}^{111} + \beta_{112i}x_{ij}^{112} + \beta_{113i}x_{ij}^{113} + \beta_{114i}x_{ij}^{114} + \beta_{115i}x_{ij}^{115} + \beta_{116i}x_{ij}^{116} + \beta_{117i}x_{ij}^{117} + \beta_{118i}x_{ij}^{118} + \beta_{119i}x_{ij}^{119} + \beta_{120i}x_{ij}^{120} + \beta_{121i}x_{ij}^{121} + \beta_{122i}x_{ij}^{122} + \beta_{123i}x_{ij}^{123} + \beta_{124i}x_{ij}^{124} + \beta_{125i}x_{ij}^{125} + \beta_{126i}x_{ij}^{126} + \beta_{127i}x_{ij}^{127} + \beta_{128i}x_{ij}^{128} + \beta_{129i}x_{ij}^{129} + \beta_{130i}x_{ij}^{130} + \beta_{131i}x_{ij}^{131} + \beta_{132i}x_{ij}^{132} + \beta_{133i}x_{ij}^{133} + \beta_{134i}x_{ij}^{134} + \beta_{135i}x_{ij}^{135} + \beta_{136i}x_{ij}^{136} + \beta_{137i}x_{ij}^{137} + \beta_{138i}x_{ij}^{138} + \beta_{139i}x_{ij}^{139} + \beta_{140i}x_{ij}^{140} + \beta_{141i}x_{ij}^{141} + \beta_{142i}x_{ij}^{142} + \beta_{143i}x_{ij}^{143} + \beta_{144i}x_{ij}^{144} + \beta_{145i}x_{ij}^{145} + \beta_{146i}x_{ij}^{146} + \beta_{147i}x_{ij}^{147} + \beta_{148i}x_{ij}^{148} + \beta_{149i}x_{ij}^{149} + \beta_{150i}x_{ij}^{150} + \beta_{151i}x_{ij}^{151} + \beta_{152i}x_{ij}^{152} + \beta_{153i}x_{ij}^{153} + \beta_{154i}x_{ij}^{154} + \beta_{155i}x_{ij}^{155} + \beta_{156i}x_{ij}^{156} + \beta_{157i}x_{ij}^{157} + \beta_{158i}x_{ij}^{158} + \beta_{159i}x_{ij}^{159} + \beta_{160i}x_{ij}^{160} + \beta_{161i}x_{ij}^{161} + \beta_{162i}x_{ij}^{162} + \beta_{163i}x_{ij}^{163} + \beta_{164i}x_{ij}^{164} + \beta_{165i}x_{ij}^{165} + \beta_{166i}x_{ij}^{166} + \beta_{167i}x_{ij}^{167} + \beta_{168i}x_{ij}^{168} + \beta_{169i}x_{ij}^{169} + \beta_{170i}x_{ij}^{170} + \beta_{171i}x_{ij}^{171} + \beta_{172i}x_{ij}^{172} + \beta_{173i}x_{ij}^{173} + \beta_{174i}x_{ij}^{174} + \beta_{175i}x_{ij}^{175} + \beta_{176i}x_{ij}^{176} + \beta_{177i}x_{ij}^{177} + \beta_{178i}x_{ij}^{178} + \beta_{179i}x_{ij}^{179} + \beta_{180i}x_{ij}^{180} + \beta_{181i}x_{ij}^{181} + \beta_{182i}x_{ij}^{182} + \beta_{183i}x_{ij}^{183} + \beta_{184i}x_{ij}^{184} + \beta_{185i}x_{ij}^{185} + \beta_{186i}x_{ij}^{186} + \beta_{187i}x_{ij}^{187} + \beta_{188i}x_{ij}^{188} + \beta_{189i}x_{ij}^{189} + \beta_{190i}x_{ij}^{190} + \beta_{191i}x_{ij}^{191} + \beta_{192i}x_{ij}^{192} + \beta_{193i}x_{ij}^{193} + \beta_{194i}x_{ij}^{194} + \beta_{195i}x_{ij}^{195} + \beta_{196i}x_{ij}^{196} + \beta_{197i}x_{ij}^{197} + \beta_{198i}x_{ij}^{198} + \beta_{199i}x_{ij}^{199} + \beta_{200i}x_{ij}^{200} + \beta_{201i}x_{ij}^{201} + \beta_{202i}x_{ij}^{202} + \beta_{203i}x_{ij}^{203} + \beta_{204i}x_{ij}^{204} + \beta_{205i}x_{ij}^{205} + \beta_{206i}x_{ij}^{206} + \beta_{207i}x_{ij}^{207} + \beta_{208i}x_{ij}^{208} + \beta_{209i}x_{ij}^{209} + \beta_{210i}x_{ij}^{210} + \beta_{211i}x_{ij}^{211} + \beta_{212i}x_{ij}^{212} + \beta_{213i}x_{ij}^{213} + \beta_{214i}x_{ij}^{214} + \beta_{215i}x_{ij}^{215} + \beta_{216i}x_{ij}^{216} + \beta_{217i}x_{ij}^{217} + \beta_{218i}x_{ij}^{218} + \beta_{219i}x_{ij}^{219} + \beta_{220i}x_{ij}^{220} + \beta_{221i}x_{ij}^{221} + \beta_{222i}x_{ij}^{222} + \beta_{223i}x_{ij}^{223} + \beta_{224i}x_{ij}^{224} + \beta_{225i}x_{ij}^{225} + \beta_{226i}x_{ij}^{226} + \beta_{227i}x_{ij}^{227} + \beta_{228i}x_{ij}^{228} + \beta_{229i}x_{ij}^{229} + \beta_{230i}x_{ij}^{230} + \beta_{231i}x_{ij}^{231} + \beta_{232i}x_{ij}^{232} + \beta_{233i}x_{ij}^{233} + \beta_{234i}x_{ij}^{234} + \beta_{235i}x_{ij}^{235} + \beta_{236i}x_{ij}^{236} + \beta_{237i}x_{ij}^{237} + \beta_{238i}x_{ij}^{238} + \beta_{239i}x_{ij}^{239} + \beta_{240i}x_{ij}^{240} + \beta_{241i}x_{ij}^{241} + \beta_{242i}x_{ij}^{242} + \beta_{243i}x_{ij}^{243} + \beta_{244i}x_{ij}^{244} + \beta_{245i}x_{ij}^{245} + \beta_{246i}x_{ij}^{246} + \beta_{247i}x_{ij}^{247} + \beta_{248i}x_{ij}^{248} + \beta_{249i}x_{ij}^{249} + \beta_{250i}x_{ij}^{250} + \beta_{251i}x_{ij}^{251} + \beta_{252i}x_{ij}^{252} + \beta_{253i}x_{ij}^{253} + \beta_{254i}x_{ij}^{254} + \beta_{255i}x_{ij}^{255} + \beta_{256i}x_{ij}^{256} + \beta_{257i}x_{ij}^{257} + \beta_{258i}x_{ij}^{258} + \beta_{259i}x_{ij}^{259} + \beta_{260i}x_{ij}^{260} + \beta_{261i}x_{ij}^{261} + \beta_{262i}x_{ij}^{262} + \beta_{263i}x_{ij}^{263} + \beta_{264i}x_{ij}^{264} + \beta_{265i}x_{ij}^{265} + \beta_{266i}x_{ij}^{266} + \beta_{267i}x_{ij}^{267} + \beta_{268i}x_{ij}^{268} + \beta_{269i}x_{ij}^{269} + \beta_{270i}x_{ij}^{270} + \beta_{271i}x_{ij}^{271} + \beta_{272i}x_{ij}^{272} + \beta_{273i}x_{ij}^{273} + \beta_{274i}x_{ij}^{274} + \beta_{275i}x_{ij}^{275} + \beta_{276i}x_{ij}^{276} + \beta_{277i}x_{ij}^{277} + \beta_{278i}x_{ij}^{278} + \beta_{279i}x_{ij}^{279} + \beta_{280i}x_{ij}^{280} + \beta_{281i}x_{ij}^{281} + \beta_{282i}x_{ij}^{282} + \beta_{283i}x_{ij}^{283} + \beta_{284i}x_{ij}^{284} + \beta_{285i}x_{ij}^{285} + \beta_{286i}x_{ij}^{286} + \beta_{287i}x_{ij}^{287} + \beta_{288i}x_{ij}^{288} + \beta_{289i}x_{ij}^{289} + \beta_{290i}x_{ij}^{290} + \beta_{291i}x_{ij}^{291} + \beta_{292i}x_{ij}^{292} + \beta_{293i}x_{ij}^{293} + \beta_{294i}x_{ij}^{294} + \beta_{295i}x_{ij}^{295} + \beta_{296i}x_{ij}^{296} + \beta_{297i}x_{ij}^{297} + \beta_{298i}x_{ij}^{298} + \beta_{299i}x_{ij}^{299} + \beta_{300i}x_{ij}^{300} + \beta_{301i}x_{ij}^{301} + \beta_{302i}x_{ij}^{302} + \beta_{303i}x_{ij}^{303} + \beta_{304i}x_{ij}^{304} + \beta_{305i}x_{ij}^{305} + \beta_{306i}x_{ij}^{306} + \beta_{307i}x_{ij}^{307} + \beta_{308i}x_{ij}^{308} + \beta_{309i}x_{ij}^{309} + \beta_{310i}x_{ij}^{310} + \beta_{311i}x_{ij}^{311} + \beta_{312i}x_{ij}^{312} + \beta_{313i}x_{ij}^{313} + \beta_{314i}x_{ij}^{314} + \beta_{315i}x_{ij}^{315} + \beta_{316i}x_{ij}^{316} + \beta_{317i}x_{ij}^{317} + \beta_{318i}x_{ij}^{318} + \beta_{319i}x_{ij}^{319} + \beta_{320i}x_{ij}^{320} + \beta_{321i}x_{ij}^{321} + \beta_{322i}x_{ij}^{322} + \beta_{323i}x_{ij}^{323} + \beta_{324i}x_{ij}^{324} + \beta_{325i}x_{ij}^{325} + \beta_{326i}x_{ij}^{326} + \beta_{327i}x_{ij}^{327} + \beta_{328i}x_{ij}^{328} + \beta_{329i}x_{ij}^{329} + \beta_{330i}x_{ij}^{330} + \beta_{331i}x_{ij}^{331} + \beta_{332i}x_{ij}^{332} + \beta_{333i}x_{ij}^{333} + \beta_{334i}x_{ij}^{334} + \beta_{335i}x_{ij}^{335} + \beta_{336i}x_{ij}^{336} + \beta_{337i}x_{ij}^{337} + \beta_{338i}x_{ij}^{338} + \beta_{339i}x_{ij}^{339} + \beta_{340i}x_{ij}^{340} + \beta_{341i}x_{ij}^{341} + \beta_{342i}x_{ij}^{342} + \beta_{343i}x_{ij}^{343} + \beta_{344i}x_{ij}^{344} + \beta_{345i}x_{ij}^{345} + \beta_{346i}x_{ij}^{346} + \beta_{347i}x_{ij}^{347} + \beta_{348i}x_{ij}^{348} + \beta_{349i}x_{ij}^{349} + \beta_{350i}x_{ij}^{350} + \beta_{351i}x_{ij}^{351} + \beta_{352i}x_{ij}^{352} + \beta_{353i}x_{ij}^{353} + \beta_{354i}x_{ij}^{354} + \beta_{355i}x_{ij}^{355} + \beta_{356i}x_{ij}^{356} + \beta_{357i}x_{ij}^{357} + \beta_{358i}x_{ij}^{358} + \beta_{359i}x_{ij}^{359} + \beta_{360i}x_{ij}^{360} + \beta_{361i}x_{ij}^{361} + \beta_{362i}x_{ij}^{362} + \beta_{363i}x_{ij}^{363} + \beta_{364i}x_{ij}^{364} + \beta_{365i}x_{ij}^{365} + \beta_{366i}x_{ij}^{366} + \beta_{367i}x_{ij}^{367} + \beta_{368i}x_{ij}^{368} + \beta_{369i}x_{ij}^{369} + \beta_{370i}x_{ij}^{370} + \beta_{371i}x_{ij}^{371} + \beta_{372i}x_{ij}^{372} + \beta_{373i}x_{ij}^{373} + \beta_{374i}x_{ij}^{374} + \beta_{375i}x_{ij}^{375} + \beta_{376i}x_{ij}^{376} + \beta_{377i}x_{ij}^{377} + \beta_{378i}x_{ij}^{378} + \beta_{379i}x_{ij}^{379} + \beta_{380i}x_{ij}^{380} + \beta_{381i}x_{ij}^{381} + \beta_{382i}x_{ij}^{382} + \beta_{383i}x_{ij}^{383} + \beta_{384i}x_{ij}^{384} + \beta_{385i}x_{ij}^{385} + \beta_{386i}x_{ij}^{386} + \beta_{387i}x_{ij}^{387} + \beta_{388i}x_{ij}^{388} + \beta_{389i}x_{ij}^{389} + \beta_{390i}x_{ij}^{390} + \beta_{391i}x_{ij}^{391} + \beta_{392i}x_{ij}^{392} + \beta_{393i}x_{ij}^{393} + \beta_{394i}x_{ij}^{394} + \beta_{395i}x_{ij}^{395} + \beta_{396i}x_{ij}^{396} + \beta_{397i}x_{ij}^{397} + \beta_{398i}x_{ij}^{398} + \beta_{399i}x_{ij}^{399} + \beta_{400i}x_{ij}^{400} + \beta_{401i}x_{ij}^{401} + \beta_{402i}x_{ij}^{402} + \beta_{403i}x_{ij}^{403} + \beta_{404i}x_{ij}^{404} + \beta_{405i}x_{ij}^{405} + \beta_{406i}x_{ij}^{406} + \beta_{407i}x_{ij}^{407} + \beta_{408i}x_{ij}^{408} + \beta_{409i}x_{ij}^{409} + \beta_{410i}x_{ij}^{410} + \beta_{411i}x_{ij}^{411} + \beta_{412i}x_{ij}^{412} + \beta_{413i}x_{ij}^{413} + \beta_{414i}x_{ij}^{414} + \beta_{415i}x_{ij}^{415} + \beta_{416i}x_{ij}^{416} + \beta_{417i}x_{ij}^{417} + \beta_{418i}x_{ij}^{418} + \beta_{419i}x_{ij}^{419} + \beta_{420i}x_{ij}^{420} + \beta_{421i}x_{ij}^{421} + \beta_{422i}x_{ij}^{422} + \beta_{423i}x_{ij}^{423} + \beta_{424i}x_{ij}^{424} + \beta_{425i}x_{ij}^{425} + \beta_{426i}x_{ij}^{426} + \beta_{427i}x_{ij}^{427} + \beta_{428i}x_{ij}^{428} + \beta_{429i}x_{ij}^{429} + \beta_{430i}x_{ij}^{430} + \beta_{431i}x_{ij}^{431} + \beta_{432i}x_{ij}^{432} + \beta_{433i}x_{ij}^{433} + \beta_{434i}x_{ij}^{434} + \beta_{435i}x_{ij}^{435} + \beta_{436i}x_{ij}^{436} + \beta_{437i}x_{ij}^{437} + \beta_{438i}x_{ij}^{438} + \beta_{439i}x_{ij}^{439} + \beta_{440i}x_{ij}^{440} + \beta_{441i}x_{ij}^{441} + \beta_{442i}x_{ij}^{442} + \beta_{443i}x_{ij}^{443} + \beta_{444i}x_{ij}^{444} + \beta_{445i}x_{ij}^{445} + \beta_{446i}x_{ij}^{446} + \beta_{447i}x_{ij}^{447} + \beta_{448i}x_{ij}^{448} + \beta_{449i}x_{ij}^{449} + \beta_{450i}x_{ij}^{450} + \beta_{451i}x_{ij}^{451} + \beta_{452i}x_{ij}^{452} + \beta_{453i}x_{ij}^{453} + \beta_{454i}x_{ij}^{454} + \beta_{455i}x_{ij}^{455} + \beta_{456i}x_{ij}^{456} + \beta_{457i}x_{ij}^{457} + \beta_{458i}x_{ij}^{458} + \beta_{459i}x_{ij}^{459} + \beta_{460i}x_{ij}^{460} + \beta_{461i}x_{ij}^{461} + \beta_{462i}x_{ij}^{462} + \beta_{463i}x_{ij}^{463} + \beta_{464i}x_{ij}^{464} + \beta_{465i}x_{ij}^{465} + \beta_{466i}x_{ij}^{466} + \beta_{467i}x_{ij}^{467} + \beta_{468i}x_{ij}^{468} + \beta_{469i}x_{ij}^{469} + \beta_{470i}x_{ij}^{470} + \beta_{471i}x_{ij}^{471} + \beta_{472i}x_{ij}^{472} + \beta_{473i}x_{ij}^{473} + \beta_{474i}x_{ij}^{474} + \beta_{475i}x_{ij}^{475} + \beta_{476i}x_{ij}^{476} + \beta_{477i}x_{ij}^{477} + \beta_{478i}x_{ij}^{478} + \beta_{479i}x_{ij}^{479} + \beta_{480i}x_{ij}^{480} + \beta_{481i}x_{ij}^{481} + \beta_{482i}x_{ij}^{482} + \beta_{483i}x_{ij}^{483} + \beta_{484i}x_{ij}^{484} + \beta_{485i}x_{ij}^{485} + \beta_{486i}x_{ij}^{486} + \beta_{487i}x_{ij}^{487} + \beta_{488i}x_{ij}^{488} + \beta_{489i}x_{ij}^{489} + \beta_{490i}x_{ij}^{490} + \beta_{491i}x_{ij}^{491} + \beta_{492i}x_{ij}^{492} + \beta_{493i}x_{ij}^{493} + \beta_{494i}x_{ij}^{494} + \beta_{495i}x_{ij}^{495} + \beta_{496i}x_{ij}^{496} + \beta_{497i}x_{ij}^{497} + \beta_{498i}x_{ij}^{498} + \beta_{499i}x_{ij}^{499} + \beta_{500i}x_{ij}^{500} + \beta_{501i}x_{ij}^{501} + \beta_{502i}x_{ij}^{502} + \beta_{503i}x_{ij}^{503} + \beta_{504i}x_{ij}^{504} + \beta_{505i}x_{ij}^{505} + \beta_{506i}x_{ij}^{506} + \beta_{507i}x_{ij}^{507} + \beta_{508i}x_{ij}^{508} + \beta_{509i}x_{ij}^{509} + \beta_{510i}x_{ij}^{510} + \beta_{511i}x_{ij}^{511} + \beta_{512i}x_{ij}^{512} + \beta_{513i}x_{ij}^{513} + \beta_{514i}x_{ij}^{514} + \beta_{515i}x_{ij}^{515} + \beta_{516i}x_{ij}^{516} + \beta_{517i}x_{ij}^{517} + \beta_{518i}x_{ij}^{518} + \beta_{519i}x_{ij}^{519} + \beta_{520i}x_{ij}^{520} + \beta_{521i}x_{ij}^{521} + \beta_{522i}x_{ij}^{522} + \beta_{523i}x_{ij}^{523} + \beta_{524i}x_{ij}^{524} + \beta_{525i}x_{ij}^{525} + \beta_{526i}x_{ij}^{526} + \beta_{527i}x_{ij}^{527} + \beta_{528i}x_{ij}^{528} + \beta_{529i}x_{ij}^{529} + \beta_{530i}x_{ij}^{530} + \beta_{531i}x_{ij}^{531} + \beta_{532i}x_{ij}^{532} + \beta_{533i}x_{ij}^{533} + \beta_{534i}x_{ij}^{534} + \beta_{535i}x_{ij}^{535} + \beta_{536i}x_{ij}^{536} + \beta_{537i}x_{ij}^{537} + \beta_{538i}x_{ij}^{538} + \beta_{539i}x_{ij}^{539} + \beta_{540i}x_{ij}^{540} + \beta_{541i}x_{ij}^{541} + \beta_{542i}x_{ij}^{542} + \beta_{543i}x_{ij}^{543} + \beta_{544i}x_{ij}^{544} + \beta_{545i}x_{ij}^{545} + \beta_{546i}x_{ij}^{546} + \beta_{547i}x_{ij}^{547} + \beta_{548i}x_{ij}^{548} + \beta_{549i}x_{ij}^{549} + \beta_{550i}x_{ij}^{550} + \beta_{551i}x_{ij}^{551} + \beta_{552i}x_{ij}^{552} + \beta_{553i}x_{ij}^{553} + \beta_{554i}x_{ij}^{554} + \beta_{555i}x_{ij}^{555} + \beta_{556i}x_{ij}^{556} + \beta_{557i}x_{ij}^{557} + \beta_{558i}x_{ij}^{558} + \beta_{559i}x_{ij}^{559} + \beta_{560i}x_{ij}^{560} + \beta_{561i}x_{ij}^{561} + \beta_{562i}x_{ij}^{562} + \beta_{563i}x_{ij}^{563} + \beta_{564i}x_{ij}^{564} + \beta_{565i}x_{ij}^{565} + \beta_{566i}x_{ij}^{566} + \beta_{567i}x_{ij}^{567} + \beta_{568i}x_{ij}^{568} + \beta_{569i}x_{ij}^{569} + \beta_{570i}x_{ij}^{570} + \beta_{571i}x_{ij}^{571} + \beta_{572i}x_{ij}^{572} + \beta_{573i}x_{ij}^{573} + \beta_{574i}$$

5. Un modèle de régression spline

Nous abordons maintenant la question que soulèvent les

boîtes à moustache de la figure 1. Examinons plus en profondeur les données observées. Les boîtes à moustache des valeurs observées de l'IMC en fonction de l'âge montrent que l'IMC est pour ainsi dire constant de 2 à 8 ans,

puis augmente à peu près linéairement de 8 à 13 ans et enfin, augmente très lentement de 14 à 19 ans. Cette caractéristique apparemment importante n'est pas incluse dans les quatre modèles. Par conséquent, à la présente section, nous essayons d'en tirer parti à l'aide d'un modèle de régression spline.

Nous utilisons la partie 1 du modèle de sélection et pour la partie 2, nous utilisons un modèle de régression joint-point. De façon générique, en posant que $c^+ = 0$ si $c \leq 0$ et $c^+ = c$ si $c > 0$, nous prenons

$$x_{ij} = \phi_{0ij} + \phi_{1ij}(a_{ij} - 8)^+ + \phi_{2ij}a_{ij} - 13^+ + e_{ij} \quad (9)$$

où, dans l'esprit de nos quatre modèles,

$$\phi_{kij} = z_{ij}^T \mathbf{a}_k + v_{kij}, \quad k = 0, 1, 2.$$

Dans (9), nous avons posé que

$$e_{ij} | \sigma_3^2 \sim \text{Normale}(0, \sigma_3^2)$$

$$x_{ij} = \phi_{0ij} + \phi_{1ij}(a_{ij} - 8)^+ + \phi_{2ij}\{a_{ij} - 13\}^+ + e_{ij},$$

nous remplaçons (9) par

rapport au modèle de sélection original. À la figure 2, nous présentons les boîtes à moustache pour DRES en fonction de l'âge. Ce diagramme est nettement meilleur que celui obtenu pour le modèle de sélection (voir la figure 1). Observons que les médianes fluctuent autour de 0 et que les variations sont faibles. Les boîtes à moustache obtenues pour 2, 3, 4, 5, 6 et 7 ans sont un peu moins variables que pour les autres âges. Nous ajustons aussi le modèle quadratique joint-point dans lequel

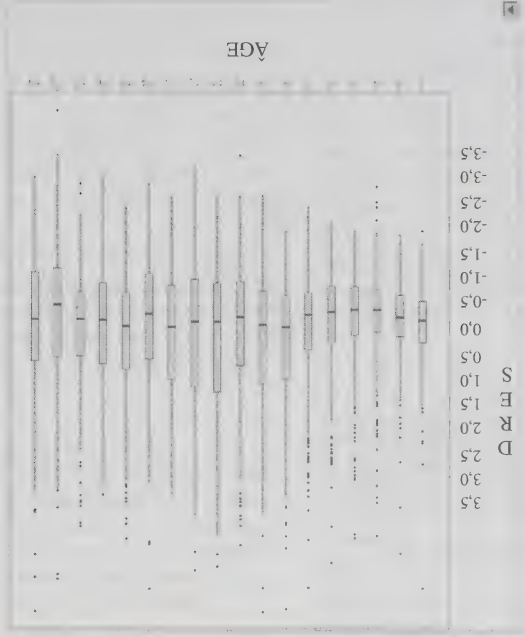


Figure 2. Boîtes à moustache des résidus de validation croisée (DRES) selon l'âge pour le modèle de régression spline

Donc, nous pouvons introduire les x_{ij} et les r_{ij} pour chaque $\Omega^{(h)}$ obtenu d'après l'algorithme MCMC à partir duquel nous obtenons M réalisations $\underline{X}_{(h)}', p'_{(h)}, h = 1, \dots, M$. Nous pouvons maintenant faire une inférence au sujet de \underline{X}_i dans (1) et de P_i dans (2).

Nous présentons les intervalles de confiance à 95 % pour la valeur moyenne en population finie (MPF) de l'IMC et la proportion en population finie (PPF) de répondants afin d'évaluer la sensibilité aux quatre modèles. Notons que nous donnons ces intervalles pour chaque domaine race selon le sexe pour chaque groupe d'âge selon le comté et que, comme ils sont fort semblables d'un domaine à l'autre, nous présentons au tableau 2 la moyenne des bornes des intervalles de confiance sur l'ensemble des comtés pour les femmes noires. Pour la MPF, les intervalles sont fort semblables d'un modèle à l'autre. Cependant, pour la PPF, ils sont fort différents. Les intervalles pour le modèle de mélange de schémas d'observation et sa version à non-réponse ignorable sont semblables, sauf pour le groupe des 2 à 4 ans, ce à quoi il faut s'attendre, puisque ces modèles expriment une régression linéaire du logarithme de la cote

Pour la PPM sous les deux modèles de mélange de schémas d'observation, les intervalles sont essentiellement les mêmes, parce que la relation de l'IMC avec l'âge, la race, le sexe et leur interaction est la même. Pour la version à non-réponse ignorable du modèle de sélection, les intervalles sont tous les mêmes sur l'âge, parce que dans la partie de ce modèle ayant trait à la réponse, l'âge et l'IMC sont tous deux ignorés. Nous notons que, pour le modèle de sélection, les intervalles ont une forme semblable à ceux obtenus pour le modèle de mélange de schémas d'observation et sa version à non-réponse ignorable. Comme l'indiquent les intervalles, la MPF et la PPF augmentent avec l'âge.

à posteriori incluse dans l'échantillonneur de Métropolis-Hastings. Les valeurs de l'IMC pour les personnes non échantillonnées doivent être prédites. Il faut souligner que nous appliquons nos modèles au logarithme de l'IMC en retenant les covariables propres à chaque individu, si bien que le logarithme de chaque valeur non échantillonnée doit être prédit, puis transformé pour le ramener à l'échelle originale. Cependant, le fait que l'on ne connaît pas l'âge, la race et le sexe pour chaque personne non échantillonnée, mais que l'on connaît le nombre de personnes dans chaque domaine âge-race-sexe pour la population américaine selon le comté réduit considérablement les calculs.

$$f(x_{ij}, r_{ij} | \mathbf{x}_{\text{obs}}, \mathbf{r}_{\text{obs}}) = \int f(x_{ij}, r_{ij} | \Omega) \pi(\Omega | \mathbf{x}_{\text{obs}}, \mathbf{r}_{\text{obs}}) d\Omega,$$

$i = 1, \dots, \ell, j = n_i + 1, \dots, N_i$. Pour le modèle de mélange de schémas d'observation, nous avons

$$f(x_{ij}, r_{ij} | \Omega) = f(x_{ij} | r_{ij}, \Omega) p(r_{ij} | \Omega)$$

et pour le modèle de sélection, nous avons

$$f(x_{ij}, r_{ij} | \Omega) = p(r_{ij} | x_{ij}, \Omega) f(x_{ij} | \Omega),$$

où Ω représente l'ensemble complet de paramètres.

Par conséquent, si nous tirons un échantillon de taille M à partir de la loi à posteriori, $\{\Omega^{(h)} : h = 1, \dots, M\}$, un estimateur de $f(x_{ij}, r_{ij} | \mathbf{x}_{\text{obs}})$ est

$$f(x_{ij}, r_{ij} | \mathbf{x}_{\text{obs}}) = M^{-1} \sum_{h=1}^M f(x_{ij}, r_{ij} | \Omega^{(h)}).$$

Tableau 2
Comparaison des quatre modèles fondée sur la moyenne des bornes des intervalles de confiance à 95 % sur l'ensemble des comtés pour la moyenne en population finie (MPF) de l'IMC et la proportion en population finie (PPF) de répondants pour les femmes noires

Modèle	2 à 4 ans	5 à 9 ans	10 à 14 ans	15 à 19 ans
SEI	MPF (14,80, 16,07)	(17,09, 18,58)	(19,63, 21,61)	(22,40, 25,19)
PPF	(0,73, 0,79)	(0,73, 0,79)	(0,73, 0,79)	(0,73, 0,79)
SE	MPF (15,55, 16,21)	(17,49, 18,36)	(19,52, 20,92)	(21,74, 23,91)
PPF	(0,66, 0,78)	(0,71, 0,81)	(0,75, 0,84)	(0,78, 0,87)
MSI	MPF (14,75, 16,10)	(17,04, 18,59)	(19,59, 21,55)	(22,42, 25,09)
PPF	(0,49, 0,70)	(0,72, 0,84)	(0,84, 0,94)	(0,90, 0,98)
MS	MPF (14,96, 15,79)	(17,16, 18,38)	(19,61, 21,45)	(22,37, 25,07)
PPF	(0,49, 0,70)	(0,73, 0,84)	(0,84, 0,94)	(0,90, 0,98)

Nota : SEI est la version à non-réponse ignorable du modèle de sélection, MSI est la version à non-réponse ignorable du modèle de mélange de schémas d'observation, MS est le modèle de mélange de schémas d'observations et SE est le modèle de sélection.

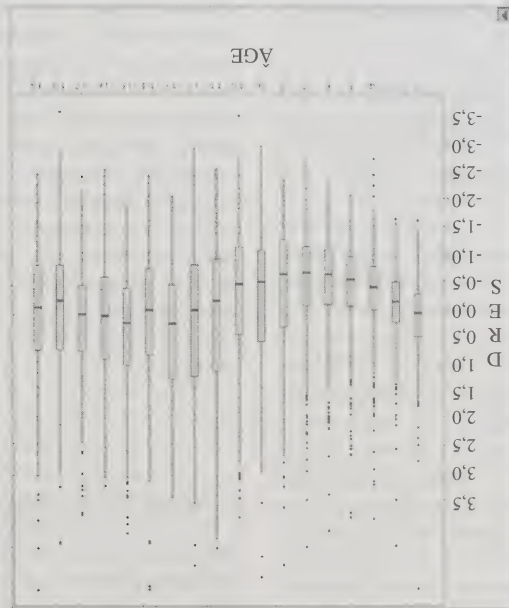
Nous étudions la relation entre l'IMC et l'âge en utilisant les intervalles de confiance à 95 % pour les paramètres du modèle de sélection. En premier lieu, nous notons que l'interaction de la race et du sexe n'est pas importante, mais que, comme il faut s'y attendre, il existe une relation importante entre l'IMC et l'âge. L'IMC augmente considérablement avec l'âge [l'intervalle de confiance à 95 % pour α_{21} est (11,89, 13,67)]. Le taux de croissance est plus faible pour les garçons de race blanche [intervalle de confiance à 95 % pour α_{22} de (-2,30, -0,19) et intervalle de confiance à 95 % pour α_{23} de (-3,03, -0,64)]. Donc, bien que l'IMC augmente avec l'âge, l'accroissement est relativement plus faible pour les garçons de race blanche. À

4.1 Estimation

À la présente section, nous analysons les données de la NHANES III sur l'IMC des enfants et des adolescents (c'est-à-dire les jeunes de 2 à 19 ans). Nous utilisons le modèle de sélection, puis, pour étudier sa sensibilité, nous comparons la prédiction sous le modèle de sélection sous non-réponse non-ignorable à celle donnée par les trois autres modèles.

4. Estimation et prédiction

Figure 1. Boîtes à moustache pour les résidus de la vérification croisée (DRES) en fonction de l'âge pour le modèle de sélection



Il faut prédire la valeur moyenne de l'IMC, ainsi que la proportion de répondants dans la population finie. Les valeurs de l'IMC pour les non-répondants échantillonnées sont obtenues au moyen de leur loi conditionnelle

4.2 Prédiction

Pour examiner plus en profondeur la question de l'ignorableté, nous traçons les boîtes à moustache (non présentées) des lois a posteriori des β_{1i} , obtenues d'après les itérations de l'algorithme de Metropolis-Hastings, selon le comté. Toutes les boîtes à moustache sont situées au-dessus de zéro, ce qui donne à penser que, pour chaque comté, le mécanisme de non-réponse est non-ignorable. En outre, il existe divers degrés de non-ignorableté. Par exemple, pour plusieurs comtés, la médiane de la boîte à moustache est proche de 1,5, tandis que pour d'autres, elle est proche de 2.

part le paramètre θ_1 , qui indique un caractère informatif (non-ignorable) important, les autres paramètres sont essentiellement sans importance. Par exemple, les intervalles de confiance à 95 % pour p_1 et p_2 sont (-0,53, 0,39) et (-0,45, 0,45), respectivement, ce qui indique qu'on pourrait utiliser un modèle plus simple (c'est-à-dire $p_1 = p_2 = 0$).

3.3 Sélection du modèle et évaluation du modèle

Nous utilisons l'approche de la perte prédictive a posteriori minimale (Gelfand et Ghosh 1998) pour sélectionner le meilleur des quatre premiers modèles. Sous l'erreur quadratique comme fonction de perte, la perte prédictive minimale a posteriori est

$$D_k = P + \frac{k+1}{k} G$$
$$P = \sum_{i=1}^n \text{Var}(x_{\text{pré}}^{(i)} | \mathbf{x}_{\text{obs}}, \mathbf{r}^{(i)})$$
$$G = \sum_{i=1}^n \{E(x_{\text{pré}}^{(i)} | \mathbf{x}_{\text{obs}}, \mathbf{r}^{(i)}) - x_{\text{obs}}^{(i)}\}^2$$

où $f(x_{\text{pré}}^{(i)} | \mathbf{x}_{\text{obs}}) = \int f(x_{\text{pré}}^{(i)} | \Omega) \pi(\Omega | \mathbf{x}_{\text{obs}}) d\Omega$ et $x_{\text{pré}}^{(i)}$ sont les valeurs prédites et Ω est l'ensemble de tous les paramètres. Cette mesure étend celle obtenue antérieurement (Laud et Ibrahim 1995) et nous prenons $k = 100$ pour établir la concordance avec cette version antérieure. Notons que, pour l'application en présence de non-réponse, nous calculons ces mesures uniquement d'après les données complètes sur l'IMC après avoir ajusté nos modèles de non-réponse.

Dans le tableau 1, nous présentons la mesure de la déviance (D_{100}) et ses composantes connexes, la qualité d'ajustement (G) et la pénalité (P), pour les quatre modèles. Si l'on se fonde sur la mesure de déviance, le modèle de sélection est nettement meilleur que les autres. Tandis que la valeur de P est à peu près la même que pour les autres modèles, celle de G est beaucoup plus petite, ce qui rend D_{100} plus petite pour le modèle de sélection. La différence entre les deux modèles de mélange de schémas d'observation est plus importante que celle entre les deux modèles de sélection. Cependant, comme nous ne disposons pas des erreurs-types, il est difficile de dire quel est le degré de signification de la différence.

Tableau 1

Comparaison des modèles de sélection et des modèles de mélange de schémas d'observation sous non-réponse ignorables et non-ignorable au moyen de la mesure de déviance

Modèle	G	P	D_{100}
SEI	135	135	270
SE	118	135	253
MSI	268	135	403
MS	204	135	339

Nota : $D_{100} = G + (100/(100+1)) P$ où G est une mesure de la qualité de l'ajustement, P est une pénalité et D_{100} est la déviance; le modèle de mélange de schémas d'observation (MS) et le modèle de sélection (SE) sont tous deux des modèles à mécanisme de réponse non-ignorable. SEI est la version à non-réponse ignorable du modèle de sélection et MSI est la version à non-réponse ignorable du modèle de mélange de schémas d'observation.

Ensuite, nous examinons les déficiences du modèle de sélection. Nous utilisons une analyse de validation croisée bayésienne pour évaluer la qualité de l'ajustement du modèle choisi (c'est-à-dire le modèle de sélection). Pour cela, nous utilisons les résidus supprimés sur les valeurs de l'IMC des répondants. Soit $(\mathbf{x}^{(i)}, \mathbf{r}^{(i)})$ le vecteur de l'ensemble des observations, sauf la (i)^e observation (x_{ij}, r_{ij}) . Alors, le (i)^e résidu supprimé est donné par

$$DRES_{ij} = \{x_{ij} - E(x_{ij} | \mathbf{x}^{(i)}, \mathbf{r}^{(i)})\} / \text{STD}(x_{ij} | \mathbf{x}^{(i)}, \mathbf{r}^{(i)})$$

Ces valeurs sont obtenues en réalisant un échantillonnage préférentiel (pondéré) sur les données de sortie de l'algorithme de Metropolis-Hastings. Nous obtenons les moments a posteriori à partir de

$$f(x_{ij} | \Omega) = \int f(x_{ij} | \Omega) \pi(\Omega | \mathbf{x}^{(i)}, \mathbf{r}^{(i)}) d\Omega$$
$$f(x_{ij} | \Omega) = f(x_{ij} | r_{ij} = 0, \Omega) p(r_{ij} = 0 | \Omega)$$
$$+ f(x_{ij} | r_{ij} = 1, \Omega) p(r_{ij} = 1 | \Omega)$$

et pour le modèle de sélection

$$f(x_{ij} | \Omega) \sim \text{Normale} \{ (\mathbf{z}_{ij}^T \mathbf{a}_1 + v_{0i}) + (\mathbf{z}_{ij}^T \mathbf{a}_2 + v_{1i}) a_{ij}, \sigma_3^2 \}$$

Nous avons également considéré l'utilisation de l'ordonnée conditionnelle a posteriori (OCP), qui est évaluée au x_{ij} observé. Cependant, ces OCP ont mené à des résultats semblables pour le repérage des valeurs extrêmes.

Nous avons tracé des boîtes à moustache (non présentées) pour DRES en fonction des quatre niveaux de sexe et des 35 comtés, ce qui nous a permis de constater que le modèle de sélection était bien ajusté. Nous avons également tracé les boîtes à moustache pour DRES en fonction de l'âge et, fait intéressant, nous avons observé une tendance. Pour le groupe des 2 à 4 ans, le modèle semble bien ajusté, tandis que pour le groupe des 5 à 9 ans, les valeurs prévues de l'IMC sont un peu élevées et pour les groupes des 10 à 14 ans et des 15 à 19 ans, la variabilité est plus importante. Nous avons examiné les boîtes à moustache pour DRES en fonction de l'âge de façon plus approfondie en traçant les boîtes à moustache pour 18 âges individuels (c'est-à-dire ceux compris entre 2 et 19 ans) (voir la figure 1). Pour les âges 11 à 19, le modèle est bien ajusté, mais pour les âges 2 à 10, il y a un problème (c'est-à-dire une courbe vers le bas dans les médianes). La même

tendance s'observe pour les trois autres modèles. Un perfectionnement supplémentaire du modèle de sélection décrit à la section 5 permet de résoudre ce problème.

Nous présentons le modèle de mélange de schémas d'observation sous non-réponse non-ignorable à l'annexe A. Nous avons inclus la race, le sexe et leur interaction dans la partie réponse du modèle, quoique cela s'avère non nécessaire. La différence entre les répondants et les non-répondants dans le modèle de mélange de schémas d'observation est que, dans la régression, l'ordonnée à l'origine varie selon le comité pour les répondants, mais non pour les non-répondants; les autres paramètres sont les mêmes. De cette façon, nous pouvons « centrer » le modèle de non-réponse non-ignorable sur le modèle de non-réponse ignorable avec une certaine variation; consulter Nandram et Choi (2002) à pour une idée comparable. Cette étape est nécessaire parce que les paramètres deviennent non identifiabiles si l'on suppose sans preuve scientifique qu'il existe une différence importante entre les répondants et les non-répondants dans le modèle de non-réponse non-ignorable. Bien que nous ayons utilisé des effets aléatoires pour faire la distinction entre les répondants et les non-répondants, les paramètres fournissant une différence systématique entre les répondants et les non-répondants dans le modèle de Rubin (1977) ne sont pas identifiables. Il convient de souligner que, dans le modèle de mélange de schémas d'observation donné en (A.4), il existe deux spécifications/schémas pour x_{ij} (i.e., $r_{ij} = 0$ et $r_{ij} = 1$), mais que dans le modèle de sélection, il n'en existe qu'une seule.

Nous montrons comment spécifier les paramètres tels que $\theta_{(0)}$, $\mathbf{a}_{(0)}^k$, $\Delta_{(0)}^k$, $k = 1, 2$ à l'annexe C. Pour obtenir une loi a priori diffuse appropriée, nous choisissons pour a une valeur telle que 0,002. Il est également possible d'utiliser une loi a priori de rétrécissement sur σ_1^{-2} et σ_2^{-2} (voir Natarajan et Kass 2000 et Daniels 1999). Néanmoins, cela n'est pas nécessaire dans le modèle hiérarchique.

L'une des propriétés intéressantes du modèle hiérarchique bayésien est qu'il introduit une corrélation entre les variables. Par exemple, dans le modèle de sélection, (4) et (7) introduisent une corrélation entre les r_{ij} et entre les x_{ij} , respectivement. Il s'agit de l'effet de mise en grappes dans les domaines. Il est possible d'obtenir ce genre d'effet directement, mais la démarche n'est pas aussi simple que dans un modèle hiérarchique. Un autre avantage du modèle hiérarchique est qu'il tient compte des variations extrinsèques entre les domaines, ce qui est intimement relié à l'effet de mise en grappes. Encore un autre avantage est que les spécifications du modèle sont robustes à des niveaux plus profonds que le processus d'échantillonnage (par exemple, l'inférence avec (5) et (8) est assez robuste à des perturbations modérées des spécifications des hyperparamètres). Nous avons observé cette robustesse empiriquement ici et dans d'autres applications.

Nous obtenons un modèle de sélection sous non-réponse ignorable en posant que $\beta_i = 0$ pour tous les comités avec

3.2 Ajustement du modèle

À la présente section, nous décrivons comment utiliser l'échantillonneur de Metropolis-Hastings pour ajuster les modèles. Nous utilisons aussi une mesure de déviance pour choisir le meilleur de nos quatre modèles. Puis, nous utilisons une analyse de validation croisée pour évaluer la qualité de l'ajustement du modèle sélectionné et, puisque les mêmes principes généraux s'appliquent aux quatre modèles, nous décrivons l'ajustement du modèle pour le modèle de sélection uniquement.

Donc, nous combinons maintenant le modèle du mécanisme de réponse et le modèle des valeurs de l'IMC pour obtenir la loi conjointe a posteriori de tous les paramètres. Les x_{ij} pour $j = r_i + 1, \dots, n_i$, $i = 1, \dots, \ell$ sont inconnus; autrement dit, ce sont des variables latentes. Nous représentons ces variables latentes par $x_{(s,m)}^{(i)}$ et les données observées par $x_{obs}^{(i)}$. En nous servant du théorème de Bayes pour combiner la fonction de vraisemblance et la loi conjointe a priori, nous obtenons la loi conjointe a posteriori qui, outre la constante de normalisation, est $p(x_{(s,m)}^{(i)}, \sigma^2, \mathbf{a}, \beta, \nu, \theta, \rho_1, \rho_2 | x_{(s,r)}^{(i)})$ et est donnée par (B.1) à l'annexe B.

La loi a posteriori (B.1) est complexe, si bien que nous utilisons des méthodes de Monte Carlo par chaîne de Markov (MCMC) pour tirer des échantillons à partir de celle-ci. Plus précisément, nous utilisons l'échantillonneur de Metropolis-Hastings (voir Chib et Greenberg 1995, pour une discussion pédagogique). Nous utilisons aussi les tracés de courbes et les diagnostics d'autocorrélation passés en revue par Cowles et Carlin (1996) pour étudier la convergence et nous suivons la proposition de Gelman, Roberts et Gilks (1996) consistant à surveiller la probabilité de saut à chaque pas de Metropolis dans notre algorithme. Durant l'exécution des calculs, le centrage des valeurs de l'IMC facilite la réalisation de la convergence (voir Gelman, Sahu et Carlin 1995). Cependant, il ne s'agit pas d'une tâche simple, car dans la régression logistique, le centrage a aussi une incidence sur la partie du modèle ayant trait à l'IMC. Nous avons obtenu un échantillon de 1 000 itérations que nous avons utilisé pour l'inférence et la vérification du modèle. En utilisant les tracés de courbes, nous avons annulé l'effet des autocorrélations, nous avons sélectionné un processus d'échantillonnage « d'apprentissage » et, pour procéder à 1 000 itérations « d'apprentissage » et, pour annuler l'effet des autocorrélations, nous avons sélectionné ensuite une itération sur dix. Nous avons obtenu cette règle par tâtonnement, durant le réglage fin des pas de Metropolis. Nous avons maintenu les probabilités de saut dans l'intervalle (0,25, 0,50); voir Gelman et coll. (1996).

aléatoirement. Il s'agit là du biais de non-réponse dont nous devons tenir compte. Il est évident que nous devons prédire la valeur de l'IMC, x_{ij} , pour a) les non-répondants dans s et b) les personnes dans ns . Donc, pour la population finie de N_i personnes, nous avons besoin d'une inférence prédictive bayésienne pour

$$\bar{X}_i = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i} \quad \text{et} \quad P_i = \frac{\sum_{j=1}^{N_i} r_{ij}}{N_i},$$

pour $i = 1, \dots, \ell$.

En posant $\bar{x}_{i(s)}^j = \sum_{r=1}^{N_i} x_{ij} / r_{ij}$, $\bar{x}_{i(ns)}^j = \sum_{r=1}^{N_i} x_{ij} / (n_i - r_i)$ et $\bar{x}_{i(n)}^j = \sum_{r=1}^{N_i} x_{ij} / N_i$, nous notons que

$$\bar{X}_i = \bar{x}_{i(s)}^j f_i^j / g_i^j + (1 - g_i^j) \bar{x}_{i(ns)}^j + (1 - f_i^j) \bar{x}_{i(n)}^j \quad (1)$$

où $f_i^j = n_i / N_i$ et $g_i^j = r_i / n_i$. Souignons que, alors que les f_i^j sont fixes en vertu du plan de sondage, les g_i^j et $\bar{x}_{i(s)}^j$ sont observés. En outre, en posant $\bar{p}_{i(s)}^j = r_i / N_i$ et $P_i^j = \sum_{r=1}^{N_i} f_i^j \bar{p}_{i(s)}^j / (N_i - n_i)$,

$$P_i^j = f_i^j \bar{p}_{i(s)}^j + (1 - f_i^j) \bar{p}_{i(ns)}^j, \quad (2)$$

$i = 1, \dots, \ell$. Nous établissons nos modèles hiérarchiques bayésiens de façon à obtenir une inférence prédictive pour des quantités comme (1) et (2) suivant le domaine.

3.1 Modèles concurrents

Nos modèles comprennent deux parties, l'une pour le mécanisme de réponse et l'autre pour la distribution de l'IMC. Ces deux parties sont reliées pour former un modèle unique sous l'hypothèse de non-réponse non-ignorable ou

de non-réponse ignorable.

Premièrement, nous décrivons le modèle de sélection. Pour la partie I de ce modèle, la réponse dépend de l'IMC

comme suit

$$r_{ij} | x_{ij}, \boldsymbol{\beta}_i \sim \text{Bernoulli} \left\{ \frac{e^{\beta_{0i} + \beta_{1i} x_{ij}}}{1 + e^{\beta_{0i} + \beta_{1i} x_{ij}}} \right\}, \quad (3)$$

$$(\beta_{0i}, \beta_{1i}) | \theta_i^0, \theta_i^1, \sigma_i^1, \sigma_i^2, p_i \sim \text{BVNormal}(\theta_i^0, \theta_i^1; \sigma_i^1, \sigma_i^2, p_i), \quad (4)$$

$$\boldsymbol{\theta} \sim N(\boldsymbol{\theta}_{(0)}, \Delta_{(0)}), \sigma_i^1, \sigma_i^2 \sim \text{Gamma}(a/2, a/2) \quad \text{et} \quad p_i \sim \text{Uniforme}(-1, 1), \quad (5)$$

où $a, \boldsymbol{\theta}_{(0)}$ et $\Delta_{(0)}$ doivent être spécifiés. Notons que, dans (5), les lois a priori sont conjointement indépendantes. L'hypothèse (3) est importante, car elle établit le lien entre la propension à répondre et les valeurs de l'IMC; les médecins pensent que les personnes qui font de l'embonpoint ou qui sont obèses ont tendance à ne pas se présenter au centre d'examen mobile pour les examens

La deuxième partie du modèle a trait à l'IMC. Le rôle de la race et du sexe étant relativement mineur. Une option consiste à poser que les valeurs de l'IMC sont

$$x_{ij} = \mu_i + \varepsilon_{ij}, \quad \mu_i = \alpha_{0i} + \alpha_{1i} a_{ij}$$

où a_{ij} dénote l'âge et $\varepsilon_{ij} \in \sigma_i^2 \sim \text{Normal}(0, \sigma_i^2)$ pour $i = 1, \dots, \ell$ et $j = 1, \dots, N_i$. En outre, il est nécessaire de comprendre la relation entre l'IMC et l'âge, la race et le sexe. Soit $z_{ij}^0 = 1$ pour une coordonnée à l'origine, $z_{ij}^1 = 1$ pour non noir et $z_{ij}^2 = 0$ pour noir, $z_{ij}^2 = 1$ pour masculin et $z_{ij}^2 = 0$ pour féminin, $z_{ij}^3 = z_{ij}^1 z_{ij}^2$ pour l'interaction entre la race et le sexe, et soit $z_{ij}^j = (z_{ij}^0, z_{ij}^1, z_{ij}^2, z_{ij}^3)$. Alors, pour la régression de l'IMC sur l'âge en corrigeant pour la race et le sexe, en posant $\boldsymbol{a}_i' = (\alpha_{0i}, \alpha_{02}, \alpha_{03}, \alpha_{04})$ et $\boldsymbol{a}_i'' = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14})$, nous prenons $\alpha_{0ij} = z_{ij}^j \boldsymbol{a}_i' + v_{0i}$ et $\alpha_{1ij} = z_{ij}^j \boldsymbol{a}_i'' + v_{1i}$ pour obtenir

$$\mu_i = (z_{ij}^j \boldsymbol{a}_i' + v_{0i}) + (z_{ij}^j \boldsymbol{a}_i'' + v_{1i}) a_{ij}$$

où v_{0i} et v_{1i} sont les effets aléatoires centrés à l'origine avec une loi normale bivariée donnée plus bas pour chaque modèle. Donc, dans la deuxième partie du modèle de sélection,

nous supposons que

$$x_{ij} = (z_{ij}^j \boldsymbol{a}_i' + v_{0i}) + (z_{ij}^j \boldsymbol{a}_i'' + v_{1i}) a_{ij} + \varepsilon_{ij} \quad \text{et} \quad \varepsilon_{ij} | \sigma_i^2 \sim \text{Normal}(0, \sigma_i^2), \quad (6)$$

$$(v_{0i}, v_{1i}) | \sigma_i^2, \sigma_i^2, p_2 \sim \text{BVNormal}(0, 0; \sigma_i^2, \sigma_i^2, p_2). \quad (7)$$

De nouveau, nous tenons compte de la mise en grappes dans les comités au moyen de l'hypothèse (7), qui est celle qui permet le « renforcement par emprunt d'information » entre les comités. Pour cette partie du modèle, nous utilisons les lois a priori

$$\boldsymbol{a}_1 \sim \text{Normal}(\boldsymbol{a}_{(0)}^1, \Delta_{(0)}^1) \quad \text{et} \quad \boldsymbol{a}_2 \sim \text{Normal}(\boldsymbol{a}_{(0)}^2, \Delta_{(0)}^2), \quad \sigma_i^2, \sigma_i^4, \sigma_i^5 \sim \text{Gamma}(a/2, a/2) \quad \text{et} \quad p_2 \sim \text{Uniforme}(-1, 1) \quad (8)$$

où $a, \boldsymbol{a}_{(0)}^k$ et $\Delta_{(0)}^k, k=1, 2$ doivent être spécifiés. Notons que, dans (8), les lois a priori sont conjointement indépendantes.

contribué une personne à l'échantillon, 22,5 %, deux personnes et 21,4 %, au moins trois personnes. Nous avons calculé le coefficient de corrélation pour les valeurs de l'IMC par appartement des membres dans les ménages (voir Rao 1973, page 199). La valeur de 0,19 obtenue indique une première approximation, nous pouvons ignorer la mise en grappes dans les ménages.

Pour les besoins de notre application, nous devons faire une inférence pour chaque domaine âge-race-sexe dans un comté. L'une des méthodes standard d'estimation sur petits domaines consiste à identifier chaque petit domaine au moyen d'un paramètre, puis à supposer qu'il existe un processus stochastique commun sur les 560 paramètres. Toutefois, à cause de la rareté des données, l'application de cette méthode n'est pas souhaitable. Donc, nous construisons nos modèles au niveau du comté et nous représenterons l'âge, la race et le sexe comme des covariables. Nous procédons à l'inférence pour chaque domaine formé par le recoupement de l'âge, de la race et du sexe dans le comté au moyen de nos modèles de régression, ce qui est un élément essentiel de notre analyse.

3. Méthode hiérarchique bayésienne

À la présente section, nous décrivons deux modèles bayésiens pour la non-réponse non-ignorable et nous déduisons deux modèles supplémentaires pour la non-réponse ignorable à titre de cas particuliers. Nous décrivons la sélection et l'évaluation du modèle pour le modèle choisi (c'est-à-dire le modèle de sélection).

Nous disposons de données provenant de $\ell = 35$ comtés et chaque comté comprend N_ℓ personnes (connues). Nous supposons qu'un échantillon probabiliste de n_ℓ personnes est tiré dans le ℓ^{e} comté. Soit s l'ensemble d'unités échantillonnées et ns l'ensemble d'unités non échantillonnées. Soit r_{ij} pour $i = 1, 2, \dots, \ell$ et $j = 1, 2, \dots, N_\ell$ l'indicateur de réponse ($r_{ij} = 1$ pour les répondants et $r_{ij} = 0$ pour les non-répondants) pour la j^{e} personne dans le i^{e} comté dans la population. En outre, soit x_{ij} le logarithme de la valeur de l'IMC. Nous avons constaté que la transformation logarithmique donnait une meilleure représentation et nous l'utilisons donc dans tout l'exposé. Il convient de souligner que les valeurs de r_{ij} et x_{ij} sont toutes observées dans l'échantillon s , mais qu'elles sont inconnues dans ns . Soit $r_\ell = \sum_{j=1}^{N_\ell} r_{ij}$ (autrement dit, r_ℓ est le nombre de personnes échantillonnées qui ont répondu dans le ℓ^{e} comté).

Par souci de commodité, nous exprimons le logarithme de l'IMC x_{ij} sous la forme $x_{i1}, x_{i2}, \dots, x_{in_\ell}, \dots, x_{iN_\ell}$ dans s et $x_{in_\ell+1}, \dots, x_{iN_\ell}$ dans ns pour le comté i . Un point important que nous tenons à souligner pour la suite est que les r_ℓ personnes ne sont pas nécessairement des répondants aléatoires provenant des n_ℓ personnes échantillonnées.

Nous étudions les données sur l'IMC pour quatre leur enfant quitte le domicile pour un examen physique. enfants est que les parents ou les mères plus âgés se sont montés extrêmement protecteurs et n'ont pas permis que 24 %. L'une des raisons de la non-réponse chez les jeunes adolescents, le taux observé de non-réponse est d'environ santé. Nous notons aussi que, pour les enfants et les plausibles que la propension à répondre soit liée à l'état de

Les données de la NHANES III sont rajustées par étapes multiples de pondération par le quotient afin de les rendre représentatives de la population; voir Mohajjar, Bell et Wakseberg (1994). Selon cette méthode d'ajustement par le quotient, la correction pour la non-réponse partielle se fait par estimation par le quotient dans la même classe d'ajustement en supposant que les distributions des répondants et des non-répondants sont identiques. Il est toutefois nécessaire de considérer d'autres méthodes d'ajustement que celle par le quotient pour traiter la non-réponse non-ignorable. Ici, nous présentons une méthode bayésienne comme option possible pour l'étude de la non-réponse dans le cas de la NHANES III.

Schaffer, Ezazli-Rice, Johnson, Khare, Little et Rubin (1996) ont entrepris de procéder à une imputation multiple complète des données de la NHANES III pour de nombreuses variables. Le but du projet était d'imputer des données pour tenir compte de la non-réponse en vue de produire plusieurs ensembles de données à grande diffusion. L'une des contraintes imposées était que la procédure utilisée pour créer les données manquantes corresponde à un mécanisme purement ignorable et que la simulation ne fournisse aucune information sur l'effet des écarts possibles par rapport au mécanisme de non-réponse ignorable. Une autre contrainte était que la procédure ne comporte pas de mise en grappes géographique. L'objectif de la présente étude est différent; nous n'avons pas l'intention de fournir des données à grande diffusion imputées. Contrairement à Schaffer et coll. (1996), nous incluons la mise en grappes au niveau du comté, bien qu'il puisse être nécessaire d'inclure la mise en grappes au niveau du ménage. Pour les données complètes, il existe 6 440 ménages. De ceux-ci, 52,1 % ont

non-répondants. Dans le cas du modèle de sélection, cette question peut être traitée au moyen de la structure hiérarchique normale. Nous considérons aussi un modèle de schémas d'observations. Le modèle de mélange de schémas d'observation est une alternative utile pour étudier la sensibilité à l'hypothèse faite dans le modèle de sélection. Pour évaluer l'hypothèse de non-réponse non-ignorable, nous considérons aussi des cas particuliers des modèles de sélection et de mélange de schémas d'observation afin d'obtenir deux modèles de non-réponse ignorable. Nous constatons qu'un cinquième modèle est nécessaire, dans lequel nous étendons notre modèle de sélection à un modèle de régression spline pour tenir compte de la relation dynamique entre l'IMC et l'âge.

Nandram, Han et Choi (2002) ont mis au point une méthode pour analyser les données sur l'IMC selon l'âge, la race et le sexe quand l'IMC est classé en trois intervalles. Cette méthode représente une extension multinomiale de l'analyse de données binaires sous non-réponse non-ignorable de Stasny (1991). Cette méthode s'applique généralement à n'importe quel nombre de cellules dans plusieurs régions (les comtés dans notre application). Nandram et Choi (2002 a, b) considèrent d'autres extensions des travaux de Stasny portant sur les données binaires (c'est-à-dire, les données provenant de la National Health Interview Survey et de la National Crime Survey). Ici, nous ne catégorisons pas les valeurs de l'IMC, mais nous les traitons plutôt, comme il se doit, comme des valeurs continues. Les quantités d'intérêt sont l'IMC moyen en population finie et la proportion de personnes qui répondent dans chaque domaine formé par l'âge, la race, le sexe et le comté.

2. Données de la NHANES III

Le plan de sondage est un plan stratifié probabiliste à plusieurs degrés qui est représentatif de l'ensemble de la population civile non placée en établissement, âgée de deux mois ou plus, des États-Unis. Le nombre de personnes

National Center for Health Statistics (1992, 1994). La collecte des données de la NHANES III comprend deux volets : le premier est la sélection de l'échantillon et l'interview des membres des ménages échantillonnés en vue de recueillir les renseignements personnels à leur sujet et le second volet est l'examen physique des personnes interviewées dans un centre d'examen mobile (CEM). L'évaluation de la santé comporte des renseignements provenant de l'examen physique, des tests et des mesures faites par des techniciens, ainsi que le prélèvement d'échantillons pour analyse.

L'échantillon a été sélectionné auprès des ménages de 81 comtés des États-Unis continentaux d'octobre 1988 à septembre 1994. Toutefois, pour des raisons de confidentialité, les données finales retenues pour l'étude provenaient des 35 plus grands comtés (de 14 États) dont la population est supérieure à 500 000 habitants, pour certains groupes d'âge selon le sexe et la race. Dans le présent article, nous analysons les données à grande diffusion provenant de ces 35 comtés, les variables démographiques sont l'âge, la race et le sexe, et l'indicateur de l'état de santé d'intérêt est l'indice de masse corporelle (IMC), qui est égal au poids en kilogrammes divisé par le carré de la taille en mètres (Kuczmarski, Carroll, Flegal et Troiano 1997). Selon l'Organisation mondiale de la santé (Consultation de l'OMS sur l'obésité 2000), un adulte dont l'IMC est égal ou supérieur à 30 est obèse, la surcharge pondérale, ou embonpoint, s'étend des adultes dont l'IMC est compris dans l'intervalle [25, 30]. Pour les enfants de 1 à 6 ans et les adolescents de 7 à 19 ans, la définition de l'embonpoint et de l'obésité varie selon l'âge.

La non-réponse peut avoir lieu dans les volets interview et examen physique de l'enquête. La non-réponse à l'interview se produit lorsque les personnes échantillonnées ne participent pas à l'interview. Certaines des personnes interviewées et incluses dans le sous-échantillon pour l'évaluation de la santé n'ont pas subi l'examen physique à la maison ou au centre d'examen mobile, et ont donc manqué la totalité ou une partie des examens physiques. Ici, nous ne considérons pas le petit nombre de personnes pour lesquelles les valeurs de l'IMC et des covariables (âge, race et sexe) manquent (c'est-à-dire les cas de non-réponse totale). Par souci de simplicité et à toutes fins pratiques, il est raisonnable d'inclure toutes les personnes avec les covariables qu'elles ont déclarées (c'est-à-dire données complètes et non-réponses partielles) dans notre analyse. Cohen et Duffy (2002) font remarquer que les enquêtes sur la santé sont un bon exemple de situation où il semble

anormaux de glucose (Dietz 1998). Donc, il serait utile d'étudier l'IMC chez les enfants et chez les adolescents en appliquant des méthodes capables de fournir une correction appropriée pour la non-réponse et une meilleure mesure de précision.

En représentant par x les covariables et par y la variable

de réponse, Rubin (1987) et Little et Rubin (1987) décrivent trois catégories de mécanisme produisant des données manquantes. Ces catégories diffèrent selon que la probabilité de réponse a est indépendante de x et de y , b) dépend de x , mais non de y et c) dépend de y et, éventuellement, de x . Dans le cas a), les données manquent entièrement au hasard (MCAR, *missing completely at random*), dans le cas b), les données manquent au hasard (MAR, *missing at random*) et dans le cas c) les données ne manquent pas au hasard (MNAR, *missing not at random*). Les modèles construits pour les mécanismes MCAR et MAR sont appelés modèles de non-réponse ignorable, si les paramètres de la variable dépendante et de la réponse sont distincts (Rubin 1976). Les modèles pour les mécanismes MNAR de données manquantes sont appelés modèles de non-réponse non-ignorable, ou non-réponse informative.

Les modèles de non-réponse peuvent être classés de façon très générale en une catégorie de modèles de sélection et une catégorie de modèles de mélange de schémas d'observation (par exemple, voir Little et Rubin 1987). Soit $[y]$ et $[r]$ la densité de probabilité de la variable de réponse y et l'indicateur de réponse r , respectivement, avec des notations évidentes pour les lois conjointes et conditionnelles. Alors, le modèle de sélection spécifique que $[y, r] = [r|y][y]$ et le modèle de mélange de schémas d'observation spécifique que $[y, r] = [y|r][r]$. L'approche par sélection a été élaborée pour étudier les problèmes de sélection d'échantillon (par exemple, Heckman 1976 et Olsson 1980). Bien que les deux modèles aient la même loi conjointe, en pratique, on spécifie les composantes $[r|y]$ et $[y]$ pour le modèle de sélection et $[y|r]$ et $[r]$ pour le modèle de schémas d'observation. Donc, ces modèles peuvent être différents.

Par conséquent, nous utilisons deux modèles de non-réponse non-ignorable, c'est-à-dire un modèle de sélection et un modèle de mélange de schémas d'observation, pour analyser les données de la NHANES III. Nous utilisons chaque modèle dans le cadre hiérarchique bayésien pour résoudre notre problème de non-réponse non-ignorable et nous comparons les résultats afin d'évaluer leur sensibilité au choix du modèle. Dans le modèle de sélection, la propension à répondre est reliée à l'IMC uniquement, si bien que le modèle de l'IMC est linéaire en fonction de l'âge, de la race, du sexe et de l'interaction de la race et du sexe. Dans le modèle de mélange de schémas d'observation, la propension à répondre est reliée à l'âge, à la race et au

sexe (mais non à l'IMC), et le modèle de l'IMC présente deux formes linéaires étroitement liées en fonction de l'âge, de la race et du sexe et de l'interaction de la race et du sexe. Ces deux modèles tiennent pour l'ensemble de la population. Les valeurs de l'IMC des non-répondants et des personnes non échantillonnées sont prédites à partir de chaque modèle. Nous préférons le modèle de sélection, parce que nous pouvons intégrer la structure dans les données de la NHANES III et que, d'après des arguments statistiques, cela s'avère véridique.

Greenlees, Kececi et Zieschang (1982) ont élaboré un modèle de régression logistique-normale pour imputer des valeurs manquantes lorsque la probabilité de réponse dépend de la variable imputée. Ils ont appliqué le modèle à des données sur les traitements et salaires provenant de la Current Population Survey (CPS). David, Little, Samuel et Triest (1986) ont comparé la méthode hot deck appliquée aux données de la CPS et le modèle de régression logistique-normale appliqué aux données sur les traitements et salaires provenant d'un même ensemble de données et ont constaté que les résultats des deux méthodes différaient fort peu. Nous notons que le modèle de régression logistique-normale est un modèle de sélection de la non-réponse ignorable, mais qu'il ne rend pas compte de la mise en grappes. Dans le cas des données de la NHANES III, pour tenir compte de la mise en grappes dans les comités, il est naturel de commencer par le modèle logistique-normale.

Notre modèle de sélection hiérarchique bayésien possède une structure spéciale. Dans la NHANES III, la propension à répondre augmente avec l'âge (la race et le sexe jouent un rôle mineur) et les médecins pensent que les personnes obèses ont tendance à ne pas se présenter au rendez-vous pour l'examen physique. Donc, étant donné les valeurs de l'IMC, comme dans Greenlees et coll. (1982), les indicateurs de réponse suivent un modèle de régression logistique où le logarithme de la valeur de l'IMC est la covariable. À leur tour, les logarithmes des valeurs de l'IMC sont distribués selon un modèle linéaire dans lequel les covariables sont l'âge, la race et le sexe. Il s'agit de l'information la plus importante que nous intégrerons dans le modèle de sélection. En outre, contrairement à Greenlees et coll. (1982), notre modèle inclut des effets de mise en grappes pour tenir compte de l'hétérogénéité entre les comités au moyen des indicateurs de réponse et des valeurs de l'IMC. Ici, chaque comité possède son propre jeu de paramètres et il existe une distribution commune sur l'ensemble de ces jeux de paramètres. Il s'agit également d'une information a priori importante que nous devons intégrer dans le modèle, ce qui est l'une des caractéristiques intéressantes de la méthode hiérarchique bayésienne.

Dans l'approche bayésienne, la principale difficulté consiste à formuler la relation entre les répondants et les

Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable aux données de la NHANES

Balگوین Nandram et Jai Won Choi *

Résumé

Nous utilisons des modèles hiérarchiques bayésiens pour analyser les données sur l'indice de masse corporelle (IMC) des enfants et des adolescents en présence de non-réponse non-ignorable, c'est-à-dire informative, tirées de la troisième National Health and Nutrition Examination Survey (NHANES III). Notre objectif est de prédire l'IMC moyen en population finie et la proportion de répondants pour les domaines formés par l'âge, la race et le sexe (covariables dans les modèles de régression) pour chacun des 35 grands comtés, en tenant compte des non-répondants. Nous utilisons des méthodes de Monte Carlo par chaîne de Markov pour ajuster les modèles (deux modèles de sélection et deux modèles de mélange de schémas de validation croisée; nous montrons que le modèle de sélection sous non-réponse non-ignorable est le meilleur des quatre modèles. Nous montrons aussi que l'inférence au sujet de l'IMC n'est pas trop sensible au choix du modèle. Nous obtenons une amélioration en incluant une régression spline dans le modèle de sélection pour tenir compte de l'évolution de la relation entre l'IMC et l'âge.

Mots clés : Validation croisée; déviance; échantillonnage de Metropolis-Hastings; modèle de régression logistique-normale; modèle de régression spline.

1. Introduction

La National Health and Nutrition Examination Survey (NHANES III) est l'une des enquêtes utilisées par le National Center for Health Statistics (NCHS) pour évaluer la santé de la population américaine. L'une des variables de cette enquête est l'indice de masse corporelle (IMC), qui est utilisé par l'Organisation mondiale de la santé pour définir l'embonpoint et l'obésité. Sous des conditions d'ignorableté de la non-réponse, les estimateurs obtenus d'après les données de la NHANES III sont biaisés, parce que le nombre de non-répondants est élevé. Par conséquent, la question qui nous préoccupe principalement ici est qu'il faut tenir compte de la non-réponse, parce que les répondants et les non-répondants pourraient avoir des caractéristiques différentes. L'objectif de l'étude est de prédire l'IMC moyen en population finie des enfants et des adolescents, poststratifiés selon le comté pour chaque domaine formé par l'âge, la race et le sexe, et de déterminer quels ajustements sont nécessaires pour tenir compte de la non-réponse non-ignorable. Notre approche consiste à ajuster plusieurs modèles hiérarchiques bayésiens de façon à refléter le mécanisme de non-réponse.

Récemment, plusieurs articles traitant de l'embonpoint et de l'obésité ont été publiés. Dans son survol du premier plan national de lutte contre l'embonpoint et l'obésité, le Directeur du Service de santé publique des États-Unis a indiqué qu'il était nécessaire de procéder à des changements radicaux dans les écoles, les restaurants, les lieux de travail et les collectivités afin de combattre l'épidémie croissante d'embonpoint et d'obésité chez les Américains. Il a déclaré dans le rapport sur l'obésité qu'il ne s'agissait ni d'esthétique ni d'apparence, mais bien d'une question de santé. Comme l'a souligné Squires (2001), le coût total des soins de santé liés à l'embonpoint et à l'obésité sont de l'ordre de 117 milliards de dollars annuellement. Les enfants qui ont un surpoids font souvent de l'embonpoint à l'âge adulte, et chez l'adulte, l'embonpoint pause un risque pour la santé (Wright, Parker, Lamont et Craft 2001). Dans un article fort intéressant fondé sur les données de la NHANES, Ogden, Flegal, Carroll et Johnson (2002) décrivent les estimations nationales les plus récentes de la prévalence et de la tendance de l'embonpoint chez les enfants et les adolescents américains. Partant d'une analyse limitée, ils concluent qu'aux États-Unis, la prévalence de l'embonpoint chez les enfants continue de croître, particulièrement chez les adolescents américains-mexicains et noirs d'origine non hispanique. Plusieurs problèmes de santé ont été associés à l'embonpoint durant l'enfance. Un type 2 est relié à la croissance de la prévalence de l'embonpoint chez les enfants (Fagot-Campagna 2000); il en est de même des facteurs de risque de maladies cardiovasculaires, des taux élevés de cholestérol et des taux

- Barnard, J., et Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955.
- Chiu, W.F., Yucel, R.M., Zanutto, E. et Zaslavsky, A.M. (2001). Using matched substitutes to improve imputations for geographically linked databases. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.
- Kalton, G. (1983). *Compensating for Missing Survey Data*. Research Report Series, Ann Arbor, MI : Institute for Social Research.
- Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.
- Krieger, N., Williams, D. et Andmoss, N. (1997). Measuring social class in U.S. public health research: Concepts, methodologies, and guidelines. *Annual Review of Public Health*, 18, 341-378.
- Lessler, J.T., et Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*. New York : John Wiley & Sons, Inc.
- Little, R.J.A., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley & Sons, Inc.
- Rubin, D.B., et Zanutto, E. (2001). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. Dans *Survey Nonresponse*, (Eds. R. Groves, R. Little et J. Eidinge), New York : John Wiley & Sons, Inc., 389-402.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London : Chapman & Hall.
- Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Logiciel pour Windows 95/98/NT disponible à <http://www.stat.psu.edu/~jls/misoftwa.html>.

4.3 Résultats

Nous avons exécuté la procédure de simulation 2 000 fois et utilisé $m = 10$ pour les méthodes AMIN et AMINp.

Les valeurs moyennes des variables de recensement dans la population étaient $\bar{y} = (40\,642, 21,65, 9,55)^T$. Le biais moyen de l'estimateur MCC était $b^{CCM} = (-5\,405, -3,97, -1,79)^T$. Les autres résultats sont résumés au tableau 3. La méthode AMINp a produit d'importantes réductions en pourcentage du biais moyen relatif (de 95,0 % à 99,5 %). La méthode ISS a réduit plus fortement les biais

que la méthode AMIN, parce que la covariable d'apartenance (code postal) était nettement plus informative que l'ensemble des covariables de modélisation (section 3.2). Puisque le mécanisme de réponse était non ignorable (les probabilités de réponse dépendaient partiellement du revenu), les mauvais résultats de la méthode AMIN, qui ne s'appuyait pas sur l'utilisation de l'information géographique pour prédire le revenu, étaient prévisibles. Notons que la méthode AMIN est biaisée et que le biais est suffisamment important pour que, avec la taille d'échantillon considérée dans le présent article, les intervalles de confiance ne couvrent jamais les valeurs hypothétiques de

Dans le cas des méthodes AMIN et AMINp, le pourcentage d'information manquante était nettement plus faible que le pourcentage moyen de données non observées. Le pourcentage d'information manquante était plus faible pour la méthode AMINp que pour la méthode AMIN. Seule la méthode AMINp a produit des intervalles bien étalonnés avec couverture correcte. Bref, la méthode AMINp combine les meilleures caractéristiques des deux autres méthodes, à savoir une couverture proche de la couverture nominale et moins d'information manquante.

Tableau 3
Résultats des simulations^(a) : réduction du biais, couverture et fraction d'information manquante

Mesure	Moyenne	AMINp	Méthode AMIN	ISS
Réduction du biais en pourcentage $100R(b_{CCM}^*, b_{CCM}) = 0$.	95,2	99,5	44,6	95,2
EDU	95,0	40,6	83,7	80,3
POV	96,8	32,6	95,1	89,8
Couverture estimée des IC à 95 % ^(c)	EDU	94,8	0,00	65,7
POV	95,2	0,00	66,0	9,92
100×fraction d'information estimée $\hat{A}_{(d)}$	EDU	0,05	0,07	0,08
POV	0,07	0,08		

(a) Fondés sur 2 000 répétitions et $m = 10$.

(b) Par définition, $100R(b_{CCM}^*, b_{CCM}) = 0$.

(c) Le résultats pour les estimations MCC étaient tous nuls.

(d) Le pourcentage moyen de données non observées était environ 17%.

5. Conclusion

Les présents travaux prolongent ceux de Rubin et Zanutto (2001) à deux égards. Premièrement, notre méthode permet d'utiliser plus d'un cas apparié par enregistré. Nous montrons théoriquement que l'efficacité de l'imputation augmente à mesure que croît le nombre de cas appariés par enregistré. Quand le coût des cas appariés est assez faible, notre méthode offre l'option d'utiliser l'information provenant de plus d'un cas apparié par enregistré pour faciliter l'ajustement des modèles d'imputation, moyennant une dépense de traitement informationnelle négligeable. Deuxièmement, la méthode AMINp ne nécessite pas de modélisation paramétrique explicite de la ou des variances résiduelles, ce qui simplifie la tâche de modélisation (particulièrement dans le cas d'analyses avec résultats multivariés). Cette approche non paramétrique permet d'appliquer notre méthode à des ensembles de données présentant des structures de modèle complexes. Dans une étude par simulation, la méthode AMINp a produit des estimations dont le biais était réduit considérablement et des intervalles de confiance dont la couverture était correcte. Bien que nous nous soyons concentrés sur l'appariement basé sur la géographie pour compléter les données sur les variables non observées couplées géographiquement, les procédures décrites dans l'article peuvent être généralisées à d'autres variables d'appariement. Par exemple, pour imputer des variables cliniques, il serait peut-être plus approprié de procéder à l'appariement à un autre patient dans le même hôpital, s'il est probable que les caractéristiques cliniques et les traitements soient plus fortement associés à l'hôpital qu'à l'emplacement géographique de la résidence du patient.

Remerciements

La présente étude a été financée en partie par le Bureau of the Census aux termes d'un contrat avec le National Opinion Research Center and Dataometrics, Inc., et par une bourse de l'Agency for Healthcare Research and Quality (AHRQ) et du National Cancer Institute (HS09869). Les auteurs remercient John Z. Ayanian d'avoir dirigé le projet de recherche Quality of Cancer Care, Mark Allen et Robert Wolf d'avoir préparé les données, Bill Wright d'avoir appuyé la présente étude, ainsi que le rédacteur adjoint et deux examinateurs anonymes de leurs commentaires constructifs.

Bibliographie

Ayanian, J.Z., Zaslavsky, A.M., Fuchs, C.S., Guadagnoli, E., Crech, C.M., Cress, R.D., O'Connor, L.C., West, D.W., Allen, M.E., Wolf, R.E. et Wright, W.E. (2003). Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. *Journal of Clinical Oncology*, 21, 1293-1300.

4.2 Méthodes d'inférence et mesures de performance

exactes de l'application décrite à la section 3, mais plutôt de l'utiliser une population artificielle caractérisée par des lois semblables à celles de la population réelle pour illustrer le fonctionnement de notre méthode et de ses concurrentes.

Les résultats préliminaires indiquaient que la performance des méthodes AMIP et AMINP était semblable; cependant, la méthode AMINP est plus simple (surtout dans les analyses avec résultats multivariés), parce qu'elle ne nécessite pas la modélisation paramétrique explicite de la variance résiduelle. Nos simulations avaient pour but de comparer la performance de la méthode AMINP (en utilisant deux cas appariés par enregistré) à trois autres méthodes d'ajustement pour la non-réponse utilisées fréquemment.

1. Méthode des cas complets (MCC)

Les moyennes de population sont estimées d'après l'ensemble des unités géocodables d'un échantillon aléatoire.

2. Imputation simple par substitut (ISS)

Il s'agit de l'utilisation habituelle de substituts. Les variables de recensement non observées pour chaque unité non géocodable sont remplacées par les valeurs des variables de recensement d'une unité sélectionnée aléatoirement dans la même grappe. L'échantillon résultant est traité comme s'il ne contenait aucune unité non géocodable; les 800 grappes comprises dans un tel échantillon sont toutes utilisées pour estimer les moyennes de population.

3. Imputation multiple normale multivariée (MNM)

Cette méthode consiste à utiliser uniquement une unité tirée aléatoirement de chacune des grappes entièrement observées dans un échantillon aléatoire pour ajuster la régression linéaire normale multivariée

$$y^j \sim N(\beta_0^j + x^j B, \Sigma),$$

avec une loi a priori non informative sur les paramètres. Le modèle est alors utilisé pour créer m ensembles d'imputations multiples pour les variables de recensement non observées en utilisant une généralisation multivariée directe de l'algorithme donné par Rubin (1987, page 167).

Nous que la MCC ne comprend ni des covariables d'appariement ni des covariables de modélisation, que la méthode AMIN utilise uniquement les covariables de modélisation et que la méthode AMINP utilise à la fois la covariable d'appariement et les covariables de modélisation.

Les données MCC et ISS sont analysées par la méthode des données complètes habituelle qui consiste à estimer la moyenne de population à partir des données à l'aide de l'estimateur approprié pour l'échantillonnage en grappes à partir d'une population finie, y compris la correction pour population finie (Cochran 1977, chapitres 9–10). Les méthodes AMIN et AMINP produisent toutes deux m ensembles de données complètes, qui sont analysés chacun par la même méthode des données complètes utilisée pour les données MCC et ISS; les m ensembles d'estimations ponctuelles et d'estimations de la variance sont alors combinés en application de la règle de combinaison de l'imputation multiple (Rubin 1987; Schafer 1997, pages 108–110). Pour chaque simulation $i \in \{1, 2, \dots, T\}$, nous dénotons les estimations ponctuelles obtenues à partir des quatre modèles par $\hat{y}_{CC}^i(t)$, $\hat{y}_{SS}^i(t)$, $\hat{y}_{MN}^i(t)$ et $\hat{y}_{NP}^i(t)$, et les moyennes de ces quantités sur l'ensemble des simulations par \bar{y}_{CC} , \bar{y}_{SS} , \bar{y}_{MN} et \bar{y}_{NP} . L'évaluation de la performance des quatre méthodes de correction pour la non-réponse sera fondée sur trois mesures :

1. **Réduction en pourcentage du biais moyen d'un estimateur relativement au biais moyen de l'estimateur MCC.** Représentons le biais moyen d'un estimateur par b_E . Alors

$$b_E = \bar{y}_E - \bar{y},$$

où $E \in \{CC, SS, MN, NP\}$. Nous définissons la réduction en pourcentage du biais moyen d'un estimateur comparativement au biais moyen de l'estimateur MCC comme étant

$$R(b_E, b_{CC}) = \frac{|b_{CC}| - |b_E|}{|b_{CC}|},$$

où b_E est un élément de b_E et b_{CC} est l'élément correspondant dans b_{CC} . Par définition, $R(b_{CC}, b_{CC})$ est nul.

2. **Couverture estimée des intervalles de confiance à 95 % nominaux pour \bar{y} .** Les intervalles produits par les estimations MCC ou ISS ont été construits sous les lois t appropriées. Pour les intervalles associés aux estimations AMIN ou AMINP, nous avons suivi la procédure décrite dans Schafer (1997, pages 109–110) et remplacé le nombre de degrés de liberté ν par la version mise à jour de Barnard et Rubin (1999).

3. **Fraction estimée d'information manquante au sujet de \bar{y} .** Pour la méthode AMIN ainsi que pour la méthode AMINP, nous avons calculé λ , une estimation de la fraction d'information manquante au sujet de \bar{y} (voir Barnard et Rubin (1999) pour l'expression la plus récente).

multivariés avec les covariables de modélisation et le code postal. Les corrélations estimées entre les résidus étaient : $r_2 = -0,194$, $r_3 \approx -0,297$ et $r_{23} \approx 0,357$, où la « variable 1 » est le revenu médian du ménage, la « variable 2 » est le pourcentage sans diplôme d'études secondaires et la « variable 3 » est le pourcentage sous le seuil de pauvreté. Ces estimations différaient significativement de zéro, indiquant qu'il fallait utiliser les versions multivariées des méthodes décrites à la section 2.3 pour produire les imputations.

3.3 Résultats de l'imputation multiple et comparaisons

Ayarian et coll. (2003) ont utilisé l'imputation par la méthode AMINP dans l'étude des prédicteurs de l'admission d'un chemothérapeute aux patients atteints d'un

cancer du côlon et du rectum. Leur modèle comprenait trois variables indicatrices pour des fourchettes de revenu actuel, ainsi que 21 autres variables représentant les caractéristiques du patient et de l'hôpital. L'analyse de l'imputation multiple montre que l'information perdue à cause de données manquantes était systématiquement inférieure à 0,1 %, proportion nettement plus faible que la fraction d'enregistrements non géocodables (3,3 %). Comme il fallait s'y attendre, les fractions les plus importantes d'information manquante ont été relevées pour les variables de revenu. Les résultats scientifiques exposés dans Ayarian et coll. (2003) n'auraient pas variés énormément si les cas pour lesquels les données étaient incomplètes avaient été éliminés. Néanmoins, dans ce genre d'étude, chaque cas est précieux et coûteux, et sauvegarder les 3,3 % pour lesquels des données manquaient représentait une contribution à l'étude.

Aux fins de comparaison, les variances des paramètres dans l'analyse portant sur les cas complets étaient, en moyenne, supérieures de 4,0 % à celles observées dans l'analyse sous imputation multiple. Cet écart en pourcentage est proche de la fraction de cas incomplets supprimés de l'analyse. Après l'introduction des imputations générées par notre méthode dans l'analyse scientifique, la précision de l'estimation de l'effet de « région rurale » a augmenté considérablement (l'utilisation des cas pour lesquels les données étaient complètes uniquement a fait augmenter la variance de 41,6 %, à cause de la concentration des enregistrements non géocodables dans les régions rurales (21,6 % d'enregistrements ruraux, mais seulement 3,1 % d'enregistrements non ruraux sont non géocodables).

4. Une étude par simulation

Cette étude par simulation a pour but de comparer la performance de notre nouvelle méthode à celle de trois

4.1 Données simulées et mécanisme de réponse

En supposant un échantillonnage en grappes avec taille d'échantillon de 800, nous avons tiré des échantillons aléatoires contenant 800 grappes. Pour chaque échantillon aléatoire, nous avons sélectionné aléatoirement environ la moitié des 800 grappes de façon à ce qu'elles contiennent un enregistrement non géocodable dans lequel les variables de recensement étaient non observées, la probabilité que les données manquent étant fonction de la race de la personne et du revenu moyen de la grappe (code postal). Nous avons simulé l'absence des données sous un modèle logit multinomial où les résultats étaient : aucune variable non observée ($w_{i0} = 1$), y_{i1} non observée ($w_{i1} = 1$), y_{i2} non observée ($w_{i2} = 1$) et y_{i3} non observée ($w_{i3} = 1$). Plus précisément, pour chaque $i = 1, 2, \dots, I$, soit $z_{i0} = 0$ et

$$z_{im} = a + b \times I \text{ (unité } m \text{ est race blanche)} \\ + c \times (\text{revenu moyen dans le code postal } i) \quad (7)$$

où $u = 1, 2, 3$. Alors,

$$\Pr(w_{im} = 1) = \exp(z_{im}) / \sum_{j=0}^3 \exp(z_{ij}) \quad \text{pour } u = 0, 1, 2, 3. \quad (8)$$

Les résultats de cette étude par simulation ont été fondés sur des ensembles de données générés par le mécanisme susmentionné avec $a = -1$, $b = 11$ et $c = 0,0003$, afin de rendre non géocodables environ 17 % des unités dans un échantillon aléatoire, avec probabilité de géocodage associée positivement à la race blanche et à un revenu plus élevé au niveau de l'État. La tâche consistait à utiliser l'échantillon aléatoire pour estimer \bar{y} , c'est-à-dire les valeurs moyennes de la population (1 696 grappes).

Nous avons établi les conditions de simulation décrites plus haut pour produire un test rigoureux de la méthode et des autres options en exagérant l'effet des données non observées et en faisant en sorte que l'absence de données soit fortement liée aux caractéristiques de l'individu et de la région. Nous n'essayions pas de simuler les conditions

3.2 Diagnostic préliminaire

Un test diagnostique simple de l'utilité des covariables d'appariement consiste à comparer les valeurs de R^2 corrigé pour les modèles de régression prédisant les trois variables de recensement contenant uniquement les covariables d'appariement et les modèles contenant les covariables d'appariement et les modèles contenant uniquement les covariables de modélisation, les modèles contenant les huit covariables de modélisation, les modèles contenant uniquement les huit covariables de modélisation et les modèles contenant les huit covariables de modélisation et le code postal. La valeur de R^2 corrigé est plus élevée pour les modèles contenant à la fois les covariables de modélisation et le code postal que pour les modèles correspondants ne contenant qu'un des deux types de covariable. Notre méthode d'imputation utilise l'information provenant des covariables d'appariement et de modélisation et, donc, devrait en principe donner de meilleurs résultats que les méthodes utilisant uniquement les covariables d'appariement ou de modélisation (comme le montre l'étude par simulation décrite à la section 4). Bien que la contribution des variables de modélisation à R^2 soit assez modeste, il est important de les inclure dans le modèle afin d'éliminer les biais systématiques et de représenter correctement les relations qui pourraient être importantes dans les modèles scientifiques.

Tableau 2 R^2 corrigé pour divers modèles de régression

Covariables de modélisation et d'appariement	Covariables de modélisation uniquement	Covariables de modélisation et d'appariement uniquement (code postal)
Revenu médian du ménage (INC)	0,091	0,453
Pourcentage sans diplôme d'études secondaires (EDU)	0,115	0,452
Pourcentage (POV)	0,047	0,327
Nombre de degrés de liberté du modèle ^(a)	26 ^(b)	1 133
Taille de l'échantillon	8 480	8 480
Nombre de degrés de liberté du résidu	8 454	7 347

(a) Avec l'ordonnée à l'origine.
(b) Les covariables de modélisation sont l'âge, le sexe (2 niveaux), la race (6 niveaux), l'état matrimonial (6 niveaux), le stade du cancer (6 niveaux), la chimiothérapie (2 niveaux), le type de cancer et la radiothérapie (3 niveaux), et la catégorie d'agrement octroyé par l'American College of Surgeons en 1999 à l'hôpital produisant le traitement (6 niveaux).
Pour déterminer s'il fallait utiliser un modèle multivarié, nous avons ajusté un modèle de régression à résultats

race, l'état matrimonial, le stade du cancer, le traitement par chimiothérapie, le type de cancer et le traitement par radiothérapie, et la catégorie d'agrement accordée par l'American College of Surgeons à l'hôpital produisant le traitement en 1999 (ACOS99). Ces variables sont observées pour les 10 176 enregistrements inclus dans le modèle d'imputation. (Certains de ces variables sont des prédictifs des analyses principales, mais la distinction est sans importance pour l'imputation.) Les valeurs moyennes au recensement X_1 , X_2 et X_3 sont observées dans les enregistrements géocodables, mais non dans les enregistrements non géocodables. Ces variables ont été traitées comme des variables de résultat du modèle d'imputation à la section 2.3. La structure des données est représentée au tableau 1.

Tableau 1

Structure des données utilisées pour l'imputation dans l'étude sur le cancer du côlon et du rectum

Données*	Huit covariables de recensement	Age	Sexe	... ACOS99	X_1	X_2	X_3
Non géocodable	✓	✓	...	✓	?	?	?
Premier appariement	✓	✓	...	✓	✓	✓	✓
Deuxième appariement	✓	✓	...	✓	✓	✓	✓
Géocodable	✓	✓	...	✓	✓	✓	✓
Premier appariement	✓	✓	...	✓	✓	✓	✓
Deuxième appariement	✓	✓	...	✓	✓	✓	✓
Premier appariement	✓	✓	...	✓	✓	✓	✓
Deuxième appariement	✓	✓	...	✓	✓	✓	✓

* Il existait 1 696 enregistrements pour chacun des six type de données.
✓ = observée ? = non observée

Avant d'ajuster le modèle, nous avons transformé les variables de résultat en pourcentage y_2 et y_3 à l'aide de la fonction logit mise à l'échelle :

(6)
$$\log \left(\frac{1 - (y - a)/(b - a)}{(y - a)/(b - a)} \right)$$

avec $a = -0,5$ et $b = 100,5$, de sorte qu'après les imputations, la transformation inverse avec arrondissement au nombre entier le plus proche produise des valeurs imputées comprises entre 0 et 100 inclusivement (Schafer 1999). De même, nous avons appliqué une transformation logarithmique à la variable de revenu y_1 , de sorte que les valeurs imputées de revenu ne soient pas négatives. Notons que les lois suivies par les variables transformées sont plus proches de la loi normale qu'elles ne le sont sur l'échelle originale (Schafer 1997). Pour simplifier la notation, nous redéfinissons y_1 , y_2 et y_3 comme étant les versions transformées.

renseignements personnels empêchaient les chercheurs d'utiliser l'ensemble de données complet pour la modélisation avec les codes postaux annexés (même sous forme cryptée). À titre d'illustration, nous utiliserons deux cas appariés par enregistrissement dans les analyses qui suivent.

3. Application : Étude sur le cancer du côlon et du rectum

La base de données sur le cancer du côlon et du rectum contient, en tout, 50 740 enregistrés patients, dont environ 3,3 % ne sont pas géocodables. Parmi ceux-ci, environ la moitié contiennent une adresse de cas postale (souvent dans une région rurale) et les autres une adresse tapée incorrectement ou une adresse appartenant à une région nouvellement développée qui ne figure pas encore dans la base de données sur les adresses. Dans le cadre d'une étude des prédicteurs de l'administration d'une chimiothérapie aux patients atteints d'un cancer du côlon et du rectum, les chercheurs ont estimé que les trois moyennes de groupe d'ilots de recensement qui suivent seraient des variables contextuelles utiles :

$$X_1 = \text{revenu médian du ménage};$$

$$X_2 = \text{pourcentage ne possédant pas de diplôme d'études secondaires};$$

$$X_3 = \text{pourcentage sous le seuil de pauvreté}.$$

Ces variables ont été observées dans les enregistrés géocodables, mais non observées dans les enregistrés non géocodables. La tâche consistait à produire des imputations multiples pour les variables de recensement non observées à l'aide des méthodes décrites à la section 2.

Chacune des moyennes de groupe d'ilots était publiée dans les données du recensement pour six groupes ethniques et les analyses scientifiques ont porté uniquement sur l'ensemble des moyennes de groupe d'ilots correspondant à la race/ethnicité de chaque patient. Pour les imputations utilisées dans Ayanian et coll. (2003), nous avons par conséquent ajusté six modèles distincts pour imputer les $18(6 \times 3)$ valeurs pour chaque cas non géocodable, puis nous avons sélectionné les trois variables pertinentes pour chaque patient; les lois conjointes pour divers groupes raciaux/ethniques n'étaient pas importantes, parce que chaque imputation ne comportait de valeurs que pour un seul groupe. Une autre solution aurait été d'utiliser la race comme variable d'appariement, mais cela nous aurait obligé à rechercher des appariements à une distance géographique nettement plus grande, ce qui aurait affaibli la valeur prédictive de l'appariement géographique.

Aux fins de l'exposé, nous supposons donc que nous disposons uniquement de la moyenne de groupe d'ilots

correspondant à la race pour chaque répondant, et que nous n'avons pas accès aux moyennes correspondant aux cinq autres races qui sont disponibles simultanément dans les données du recensement. Cette situation est plus typique des données qui seraient recueillies directement auprès du répondant, où la variable de race proprement dite (en tant que variable de modélisation) est relativement prédictive, parce que les données sur le revenu de personnes de races différentes reflètent les différences de revenu associées à la race.

3.1 Appariement et l'ensemble de données

L'adresse de plus de 90 % des enregistrés non géocodables contient le code postal. Par conséquent, nous avons choisi ce dernier comme covariable d'appariement. Un diagnostic simple de son utilité figure à la section 3.2. La série numérique de codes postaux ne correspond pas toujours aux relations de distance entre les quartiers. Par exemple, Cambridge, Massachusetts, possède un bureau de poste 02138 qui utilise aussi le code postal 02238 pour les boîtes aux lettres et à Boston, située tout près de là, il existe un code postal 02215 qui a été pris à la région 02115. Au lieu d'utiliser la série numérique de codes postaux, nous avons calculé les distances entre les codes postaux d'après les latitudes et longitudes du bureau de poste principal correspondant, sous l'hypothèse que deux codes postaux étaient les plus proches l'un de l'autre si leurs bureaux de poste principaux étaient les plus proches l'un de l'autre.

La base de données sur le cancer du côlon et du rectum contient 1 696 enregistrés non géocodables. Nous avons sélectionné le même nombre ($n^* = 1 696$) d'enregistrés géocodables aléatoirement à partir de la même base de données. Pour chacun de ces 3 392 enregistrés, nous avons sélectionné aléatoirement deux cas géocodables appartenés à partir du code postal de l'enregistré en question ou (au besoin) de codes postaux voisins. Nous avons obtenu ainsi un ensemble de données contenant $3\,392 \times 3 = 10\,176$ enregistrés. Notons qu'il était commun de choisir n^* , parce que les données étaient gratuites. En général, le choix de n^* pourrait avoir une incidence sur le coût total ainsi que sur la précision des estimations. Tant les enregistrés géocodables que les cas appariés sélectivement aléatoirement correspondaient à des données intra-échantillon et ont donc été retenus dans les analyses menées par Ayanian et coll. (2003). Nous avons demandé au Registre du cancer de nous fournir uniquement ces cas, pour des raisons de confidentialité, nous ne pouvions procéder nous-mêmes à l'appariement aux données que nous possédions (pour les mêmes cas).

Les covariables de modélisation que nous avons utilisées dans le modèle d'imputation étaient les huit variables d'enregistrés administratifs, à savoir l'âge, le sexe, la

Dénotons les résultats pour ces cas appariés par le vecteur $Y_i = (y_{i1}, \dots, y_{ik})^T$ et les caractéristiques correspondantes par la matrice $X_i = (x_{i1}, \dots, x_{ik})^T$. Avec une loi a priori uniforme pour δ_i , la loi a posteriori de $\delta_i | y_i, X_i, \beta$ est de moyenne

$$(3) \quad \bar{y}_i - x_i^T \beta$$

et de variance σ^2/K_i , où $\bar{y}_i = \sum_{k=1}^{K_i} y_{ik}/K_i$ et $x_i = \sum_{k=1}^{K_i} x_{ik}/K_i$. Donc, la loi prédictive pour $y_{i0} | y_i, X_i, x_{i0}, \beta$ est de moyenne

$$(4) \quad \bar{y}_i + (x_i^{i0} - x_i^T) \beta$$

1. tirer des cas appariés pour les enregistrements non géocodables et pour certains enregistrements géocodables échantillonnés aléatoirement;
2. utiliser les enregistrements géocodables échantillonnés et leurs cas appariés pour ajuster l'équation (1), où les δ_i sont traités comme des effets fixes et sauvegarder les résidus;
3. répéter m (habituellement de 5 à 10) fois les étapes suivantes :
- a) tirer σ^2 à partir de sa loi a posteriori, puis β sachant le tirage de σ^2 ;
- b) pour chaque enregistrement non géocodable, traiter la somme du vecteur des moyennes prédictives obtenu à partir de l'équation (4) et d'un vecteur de résidu tiré en utilisant soit AMNP soit AMNP en tant que réalisation du vecteur non observé de variables contextuelles.

2.4 Efficacité

L'efficacité d'une imputation dépend du nombre de cas appariés utilisés. Soit V_K la variance prédictive d'un modèle d'imputation, où K cas appariés par enregistrement sont utilisés. Pour le modèle de la section 2.3, $V_K = (1 + 1/K) \sigma^2$. Définissons l'efficacité comme étant

$$(5) \quad E_K = \frac{V_\infty}{V_K} = \frac{\sigma^2}{\sigma^2 (1 + 1/K)} = \frac{K}{K + 1},$$

pour tout nombre entier positif K . L'efficacité augmente parallèlement au nombre de cas appariés par enregistrement; par exemple, $E_2 \approx 0,67, E_4 \approx 0,8, E_{10} \approx 0,91$ et $E_{20} \approx 0,95$. En théorie, il peut exister autant de cas appariés par enregistrement que le permettent les ressources disponibles. En pratique, le nombre de cas appariés utilisés dépend souvent du coût de l'obtention de ces cas et de celui du traitement informatique requis pour ajuster le modèle. Dans le cas de notre méthode, le coût du traitement informatique nécessaire pour chaque cas apparié supplémentaire est négligeable. Dans l'étude du traitement du cancer du côlon et du rectum, l'obtention des cas appariés était gratuite, mais il était essentiel de pouvoir procéder à l'imputation sur un nombre limité de cas appariés, parce que des contraintes de protection des

Nous pouvons obtenir le résidu par modélisation ou par échantillonnage. La modélisation comprend l'estimation de σ^2 en utilisant la variance résiduelle de l'équation (1) et en tirant le résidu sous une loi normale unitaire [voir Rubin et Zanutto (2001) pour le cas particulier où un seul cas apparié a été obtenu pour chaque enregistrement] ou sous une autre loi paramétrique. Nous donnons à cette approche le nom de **AMNP paramétrique** (AMNP). Une autre option consiste à échantillonner aléatoirement un résidu de régression à partir de toute région j dont les résidus pourraient être considérés comme permutable avec ceux provenant de la région i (Rubin 1987, pages 166–168). Consulter aussi Lessler et

2.2 Appariement hors échantillon et

intra-échantillon

Les cas appariés peuvent être obtenus à partir de données hors échantillon ou de données intra-échantillon. Dans l'approche de Rubin et Zanutto, les substitués appariés sont obtenus à partir de données hors échantillon, après avoir décelé l'absence de données. La description de ces auteurs met l'accent sur le fait que les substitués appariés doivent être éliminés après l'imputation, puisque le fait d'inclure ces cas supplémentaires dans l'inférence modifierait le plan de sondage par ajout de cas supplémentaires dans les « îlots » qui contiennent des données non observées. Les cas appariés sont considérés comme provenant de données intra-échantillon s'ils sont obtenus à partir de la base de données disponible avant l'imputation ou même avant de déterminer quels enregistrements de la base de données contiennent des variables non observées. En ce qui concerne les objectifs globaux d'inférence, au lieu d'être des cas supplémentaires, ces cas appariés font partie de la série originale de données et, par conséquent, seront inclus dans les analyses scientifiques.

En supposant que l'appariement est intra-échantillon, nous traitons les enregistrements non géocodables comme des non-répondants et les enregistrements géocodables, comme des répondants. Pour chaque enregistrement non géocodable, nous choisissons aléatoirement un nombre donné de cas appariés à partir d'un groupe d'enregistrements géocodables dans la même petite région géographique (par exemple, code postal, c'est-à-dire un code de livraison postale qui, aux États-Unis, représente habituellement une région desservie par un bureau de poste principal unique). De façon semblable, nous choisissons le même nombre de cas appariés pour chaque enregistrement géocodable échantillonné aléatoirement [voir Rubin et Zanutto (2001) pour des recommandations sur la taille d'un échantillon de ce genre comparativement au nombre total d'enregistrements non géocodables dans un ensemble de données particulier]. Si un nombre de cas appariés plus grand que celui disponible dans la même petite région était nécessaire, le groupe de sélection serait étendu aux régions géographiques « les plus proches » jusqu'à l'obtention du nombre requis.

Dans l'étude du traitement du cancer du colon et du rectum, tous les cas appariés provenaient de la même base de données sur le cancer. En général, les cas appariés ne doivent pas nécessairement être tirés de la même population que celle dont proviennent les non-répondants et les répondants. Par exemple, pour les enregistrements de cas de cancer du colon et du rectum, on peut obtenir les cas appariés à partir d'une population générale de cancéreux, puis ajuster un modèle pour tenir compte des différences systématiques. Notons que, si les cas appariés proviennent

2.3 Modélisation et imputation multiple

ment cohérentes.

D'une population fort semblable, il est possible de construire des modèles plus robustes contenant un plus grand nombre de covariables. Dans notre exemple, puisque nous utilisons d'autres patients atteints de la même forme de cancer, les relations en ce qui concerne les variables de processus théorapénique et les variables de résultat sont vraisemblablement vraies vérifiées dans la population la relation suivante

$$y_{ik} = x_{ik}^T \beta + \delta_i + \varepsilon_{ik} \quad (1)$$

où l'indice i désigne la petite région géographique et l'indice k , l'unité dans la région, et y_{ik} et x_{ik} sont, respectivement, la réponse et les caractéristiques de la k^{e} unité dans la région géographique i . Ce modèle comprend une prévision par régression $x_{ik}^T \beta$, un effet de petite région δ_i , et un résidu particulier à l'unité ε_{ik} . Nous supposons que ε_{ik} suit une loi F^* de moyenne nulle et de variance σ^2 . Notons que ce développement se généralise directement à une réponse y_{ik} multivariée.

Nous étendons la méthode de Rubin et Zanutto de façon à permettre plus d'un appariement dans la même petite région, parce que l'obtention de plusieurs appariements dans les petites régions est possible (souvent commode et peu coûteuse) dans le cas des données de recensement ou de grands ensembles de données administratives. L'hypothèse d'un seul appariement émise par Rubin et Zanutto est appropriée dans le cas de la collecte de données d'enquête nécessitant du travail supplémentaire sur le terrain pour chaque appariement.

Nous estimons les coefficients de régression de l'équation (1) à l'aide d'une série d'observations associées à deux enregistrements ou plus par petite région en vue d'ajuster le modèle de régression dans lequel les δ_i sont traités comme des effets fixes. S'il n'existe que deux cas par région, on peut estimer β à partir de la régression dans la

$$(y_{i1} - y_{i2}) = (x_{i1}^T - x_{i2}^T) \beta + (\varepsilon_{i1} - \varepsilon_{i2}) \quad (2)$$

où l'effet de petite région s'élimine. Les résidus de cette régression suivent une loi symétrique de variance $2\sigma^2$.

En supposant pour le moment que nous avons procédé à un tirage à partir de la loi a posteriori de β , nous exécutons le reste de l'analyse sachant ce tirage. Supposons maintenant que nous voulions faire une imputation pour une nouvelle unité (dont l'indice est $k = 0$) dans la région i , et que nous ayons obtenu $K_1 \geq 1$ cas appariés pour cette unité.

Rubin et Zanutto (2001) utilisent l'expression « substitut apparié » au lieu de « cas apparié » et proposent un modèle d'imputation paramétrique utilisant un seul substitut apparié par enregistré. Les résultats des analyses réalisées au moyen de leur modèle ont été comparés à ceux obtenus par d'autres méthodes analytiques dans le cadre d'une grande étude par simulation, mais la méthode n'a pas été appliquée à des données réelles. Nous étendons la méthode de Rubin et Zanutto (1) en permettant l'utilisation d'information provenant de plus d'un cas apparié par enregistré (et 2) en utilisant une loi empirique plutôt que paramétrique pour les résidus.

La présente étude a été motivée par la nécessité de procéder à des imputations multiples pour les variables partiellement observées dans l'étude des profils de traitement chez les personnes atteintes d'un cancer du côlon et du rectum. O'Connor, West, Allen, Wolf et Wright (2003) ont analysé un ensemble de données comprenant des imputations générées par notre méthode, en faisant référence à Rubin et Zanutto (2001) et à une version provisoire du présent article qui a paru dans un recueil de comptes rendus (Chiu, Yucel, Zanutto et Zaslavsky 2001). Le présent article est la première publication complète de notre méthode et le premier rapport jamais publié décrivant une application de la méthode de Rubin et Zanutto à des données réelles.

La présentation de la suite de l'article est la suivante. À la section 2, nous résumons la méthode de Rubin et Zanutto, puis donnons une description générale de la nôtre. À la section 3, nous décrivons dans les grandes lignes l'application de notre méthode à l'étude du traitement du cancer du côlon et du rectum. À la section 4, nous illustrons, par une étude par simulation, les résultats de notre méthode comparativement à trois méthodes de correction pour la non-réponse utilisées fréquemment.

2. Méthode d'imputation

Nous commençons par résumer la méthode de Rubin et Zanutto, puis nous donnons une description générale de notre méthode comprenant une discussion de l'appariement hors échantillon par opposition à l'appariement intra-échantillon, les détails de la modélisation et des tâches d'imputation multiple, ainsi qu'une analyse de l'efficacité en fonction du nombre de cas appariés utilisés.

2.1 Appariement, modélisation et imputation multiple

Rubin et Zanutto (2001) ont proposé une méthode appelée « appariement, modélisation et imputation multiple » (AMI) qui consiste à utiliser des substituts appariés pour

produire des imputations multiples pour les non-répondants aux enquêtes par sondage, sans exiger que les substituts soient des remplacements parfaits des non-répondants. Les substituts appariés sont des unités qui répondent à l'enquête choisies de façon à ce qu'elles coïncident avec les non-répondants sur une ou plusieurs « covariables d'appariement », c'est-à-dire des variables pour lesquelles les données sont disponibles avant le sondage et qui sont communes pour l'appariement, mais pas forcément pour la modélisation. À cause de l'appariement, les non-répondants et leurs substituts peuvent posséder des valeurs en commun pour les « covariables de terrain », c'est-à-dire des variables qui ne sont observées qu'implicitement et ne sont, par conséquent, pas disponibles pour l'analyse des données. Les « covariables de modélisation » sont des variables qui peuvent être incluses dans les modèles statistiques pour tenir compte des différences observées entre les non-répondants et leurs substituts, mais qui pourraient ne pas être disponibles pour les covariables d'âge et d'adresse pour toutes les unités d'une population avant l'échantillonnage. L'obtention de substituts coïncidant avec les non-répondants en ce qui concerne l'âge ainsi que l'adresse pourrait être difficile. Notre solution consiste à fonder l'appariement uniquement sur l'adresse (par exemple, choisir un voisin comme substitut) et à corriger les différences systématiques d'âge entre les non-répondants et les substituts appariés par modélisation statistique. Si des ménages voisins étaient choisis comme substituts appariés des ménages non-répondants, il se pourrait que les substituts et les non-répondants vivent dans le même contexte socioéconomique (par exemple, taux de criminalité, accès aux moyens de transport en commun, etc.) même si l'on n'a pas enregistré ces caractéristiques. Dans le présent exemple, l'adresse est une covariable d'appariement, l'âge est une covariable de modélisation et les caractéristiques socioéconomiques contextuelles sont les covariables de terrain.

Brèvement, la méthode AMI consiste à i) choisir des substituts appariés pour les non-répondants et certains répondeurs d'après des covariables d'appariement, ii) utiliser des covariables de modélisation pour ajuster un modèle d'estimation des différences systématiques de réponse entre les pères répondant-substitut, iii) procéder à l'imputation multiple des valeurs non observées à l'aide du modèle obtenu en (iii) sous l'hypothèse que la même relation est vérifiée entre les pères non-répondant-substitut et iv) éliminer tous les substituts appariés après l'imputation.

Utilisation de substituts appariés pour améliorer les imputations dans les bases de données couplées géographiquement

Wai Fung Chiu, Recai M. Yucei, Elaine Zanutto et Alan M. Zaslavsky¹

Résumé

Lorsqu'on couple géographiquement les enregistrements d'une base de données administratives à des groupes d'îlots de recensement, les caractéristiques locales tirées du recensement peuvent être utilisées comme variables contextuelles susceptibles de compléter utilement les variables qui ne peuvent être observées directement à partir des dossiers administratifs. Les bases de données contenant souvent des enregistrements dont les renseignements sur l'adresse ne suffisent pas pour le couplage géographique avec des groupes d'îlots de recensement; par conséquent, les variables contextuelles pour ces enregistrements ne sont pas observées. Nous proposons une nouvelle méthode qui consiste à utiliser l'information provenant des « cas appariés » et des modèles de régression multivariable pour créer des imputations multiples pour les variables non observées. Notre méthode donne de meilleurs résultats que d'autres dans les études par simulation au moyen de données du recensement et a été appliquée à un ensemble de données choisi pour étudier les profils de traitement des personnes atteintes d'un cancer du côlon et du rectum.

Mots clés : Non-réponse totale; imputation multiple; variables contextuelles; substituts appariés; dossiers administratifs.

1. Introduction

Afin d'étudier les profils de traitement des personnes atteintes d'un cancer du côlon et du rectum, le revenu et le niveau de scolarité sont des variables qu'il est souhaitable d'utiliser pour construire des modèles statistiques pertinents du point de vue scientifique. Malheureusement, les données individuelles sur ces variables ne peuvent être extraites directement des bases de données des registres du cancer créées d'après les dossiers hospitaliers, qui, comme de nombreuses bases de données administratives, contiennent principalement des renseignements requis à des fins administratives. Par conséquent, on utilise les valeurs moyennes de ces variables pour de petites régions géographiques (groupe d'îlots ou secteur de recensement) comprenant le lieu de résidence du sujet comme variable indépendante afin d'estimer les effets du revenu et du niveau de scolarité. Les analyses fondées sur ce genre de « variables contextuelles » sont fréquentes en épidémiologie et en recherche sur les services de santé (Kriger, Williams et Andriamso 1997), et produisent souvent des résultats semblables, de façon générale, à ceux obtenus en se fondant sur des variables individuelles. Si l'on disposait à la fois de variables individuelles et de variables contextuelles, il serait possible de faire la distinction entre les effets des caractéristiques des individus et du contexte; dans une analyse purement contextuelle, ces effets sont confondus. Néanmoins, l'observation d'associations entre les caractéristiques socioéconomiques

contextuelles et la qualité des soins donneraient à penser qu'il existe un problème d'équité, que ces associations reflètent principalement des relations de niveau individuel ou des relations de niveau communautaire.

Dans l'étude du traitement du cancer du côlon et du rectum, on suppose que chaque variable contextuelle pour une variable par couplage géographique de l'adresse figurant dans l'enregistrement à un groupe d'îlots (ou secteur) de recensement. Un pourcentage faible, mais important, d'enregistrements patient (environ 3,3 %, soit 1 696 enregistrements) ne contiennent pas suffisamment d'information sur l'adresse pour permettre les couplages aux groupes d'îlots de recensement, ce qui rend les variables contextuelles correspondantes inobservables. Nous dirons que les tels enregistrements sont *non géocodables*, tandis que les enregistrements qui peuvent être couplés à des groupes d'îlots de recensement sont *géocodables*. Pour gérer des imputations multiples pour les variables contextuelles non observées, nous proposons une stratégie qui consiste à utiliser l'information provenant de plus d'un « cas apparié » pour faciliter la création de modèles paramétriques/non paramétriques d'imputation. Plus précisément, l'information provenant des cas appariés rend compte des effets de petite région dans le modèle d'imputation, si bien qu'il n'est pas nécessaire de les modéliser explicitement.

1. Wai Fung Chiu, Department of Statistics, Harvard University, One Oxford Street, Cambridge MA 02138. Courriel : wfchiu@post.harvard.edu; Recai M. Yucei, Department of Biostatistics and Epidemiology, 408 Arnold House, School of Public Health and Health Sciences, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304. Courriel : yucei@schoph.umass.edu; Elaine Zanutto, The Wharton School, University of Pennsylvania, 466 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia PA 19104. Courriel : zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA 02115. Courriel : zaslavsky@hcp.med.harvard.edu.

Citrinani, A., Di Zio M., Luzi O. et Seebert, A.C. (2000). The new integrated data editing procedure for the Italian Labour Cost survey: Measuring the effects on data of combined techniques. *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, 7-21.

De Waal, T. (2003). Résolution du problème de localisation des erreurs par la génération de sommes. *Techniques d'enquête*, 29, 1, 81-90.

Di Zio, M., Guamerà, U. et Rocci, R. (2004). A mixture of mixture models to detect unitary measure error. *Proceedings in Computational Statistics*, Verlag, Prague, August 23-28.

Di Zio, M., et Luzi, O. (2002). Combining methodologies in a data editing procedure: an experiment on the survey of Balance Sheets of Agricultural Firms. *Italian Journal of Applied Statistics*, 14, 1, 59-80.

Encyclopedias of Statistical Sciences (1999). New York: John Wiley & Sons, Inc. Mise à jour, 3, 621-629.

Eureidit (2003). *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Eureidit project*, 1, 2. A paratrait. Maintenant disponible à <http://www.csws.york.ac.uk/eureidit/>.

Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18.

Fellegi, I.P., et Holt, D. (1976). A systematic approach to edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Fraley, C., et Raftery, A. (2002). Model-Based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Granquist, L. (1995). Improving the traditional editing process. Dans *Business Survey Methods*, (Eds. B.G. Cox et D.A. Binder). Granquist, L. (1996). The new view on editing. *Revue Internationale de Statistique*, 65, 3, 381-387.

Granquist, L., et Kovar, J. (1997). Editing of survey data: How much is enough? Dans *Survey Measurement and Process Quality*, (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin), New York: John Wiley & Sons, Inc., 415-435.

Hawkins, D.M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics*, 23, 105-110.

Kovar, J.G., Mac Millan, I.H. et Whitridge, P. (1988). Overview and strategy for the generalized edit and imputation system. (mis à jour février 1991). Statistique Canada, document de travail, direction de la méthodologie, BSMMD-88-007E/F.

Laiouche, M., et Berthelot, J.M. (1992). Use of a score function to proruise and limit recontacts in business surveys. *Journal of Official Statistics*, 8, 389-400.

Lawrence, D., et McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.

McLachlan, G.J., et Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.

McLachlan G.J., et Peel D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

Anderson, T.W. (1984). *An introduction to Multivariate Statistical Analysis*. Deuxième édition. New York: John Wiley & Sons, Inc.

Azzalini, A., et Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society (B)*, 65, 367-389.

Azzalini, A., Dal Cappello, T. et Kotz, S. (2003). Log-skew-normal and log-skew-t distributions as models for family income data. *Journal of Income Distribution*, 11, 13-21.

Biernacki, C., Celeux, G. et Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gauss mixture models. *Computational Statistics & Data Analysis*, 41, 561-575.

Bibliographie

Nous remercions les examinateurs et le rédacteur associé de leurs commentaires constructifs.

Remerciements

à cause des contraintes caractéristiques de notre modèle.

nombre de paramètres reliés au vecteur de moyennes et à la matrice des covariances augmente nettement plus lentement.

variables qui peuvent être traitées simultanément. En réalité, le nombre de grappes et, donc, le nombre de paramètres de mélange π , peuvent croître exponentiellement relativement au nombre de variables, ce qui rend l'estimation des paramètres ardue. Cependant, nous mentionnerons que le nombre de paramètres reliés au vecteur de moyennes et à la matrice des covariances augmente nettement plus lentement.

Enfin, une dernière préoccupation a trait au nombre de variables qui peuvent être traitées simultanément. En réalité, le nombre de grappes et, donc, le nombre de paramètres de mélange π , peuvent croître exponentiellement relativement au nombre de variables, ce qui rend l'estimation des paramètres ardue. Cependant, nous mentionnerons que le nombre de paramètres reliés au vecteur de moyennes et à la matrice des covariances augmente nettement plus lentement.

un « mélange de modèles de mélanges » (McLachlan et Peel 2000; Di Zio, Guamerà et Rocci 2004).

la comparaison plus directe, nous remplaçons pour ces unités les valeurs incorrectes par les valeurs « vraies » et obtenons $B(X_1) = B(X_2) = 0$. Ce degré de performance particulière-mment élevé du modèle s'explique par le faible degré de superposition des grappes, comme le montre clairement la figure 7.

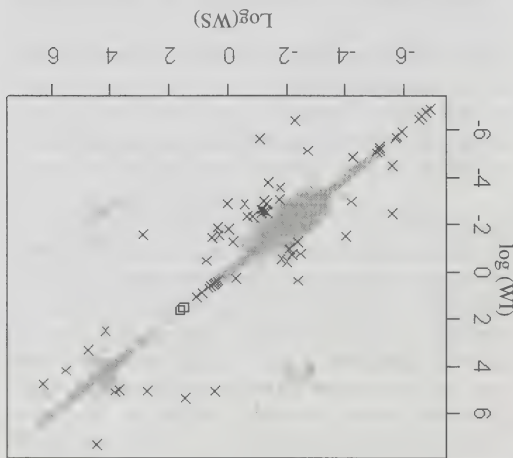


Figure 7. Diagramme de dispersion de $\log(WI)$ par rapport à $\log(WS)$. Les croix indiquent les unités critiques pour l'atypicité et les carrés, les unités critiques pour l'effet de leur erreur éventuelle.

5. Mot de la fin et futurs travaux

Dans le présent article, nous proposons un modèle de

mélanges finis pour traiter un type particulier d'erreur systématique qui entache fréquemment les données d'enquête numériques continues. L'approche proposée a les avantages, comparativement aux approches classiques, d'annoncer facilement le problème dans un contexte multivarié, d'être naturellement des diagnostics utiles pour établir l'ordre de priorité des unités douze contenant éventuellement des erreurs influentes. Cette dernière caractéristique est particulièrement importante quand la situation est critique, c'est-à-dire quand différents patrons d'erreur se superposent ou, autrement dit, quand les erreurs d'unité de mesure se situent parmi les observations plausibles. Dans ces circonstances, un examen manuel est nécessaire. Par conséquent, il est important d'optimiser la sélection des observations critiques afin de gagner du temps et d'économiser de l'argent. Tous ces avantages sont dus à l'adoption d'une approche basée sur un modèle. Par ailleurs, il est évident qu'une telle approche sous-entend des problèmes associés aux hypothèses sous-jacentes. Cependant, si l'on s'en tient aux expériences décrites dans l'article, il semble que la

Nous avons comparé ces résultats à ceux obtenus par la procédure officielle. Sur les 1 968 unités non sélectionnées pour un examen manuel, 1 911 observations sont sans erreur ou entachées d'une erreur d'unité de mesure uniquement. Pour toutes, la classification donnée par le modèle de mélanges est correcte. Parmi les 57 unités caractérisées par d'autres typologies d'erreur, 45 sont classées dans la catégorie des unités non affectées par l'erreur d'unité de mesure, tandis que 12 sont classées dans la catégorie de dernière erreur de classification peut être expliquée par l'existence d'autres erreurs systématiques (facteurs 100 et 10 000) qui ne sont pas prises en compte dans le modèle utilisé pour notre exemple.

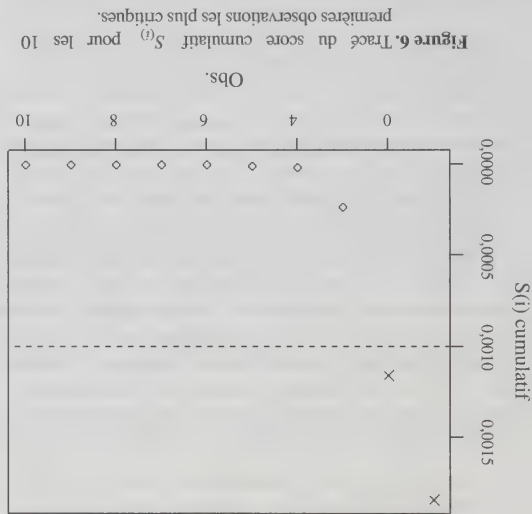


Figure 6. Tracé du score cumulé $S(i)$ pour les 10 premières observations les plus critiques.

Une autre comparaison a trait à l'estimation des totaux. Sous l'hypothèse que les valeurs sélectionnées pour un examen manuel des valeurs critiques sont groupées convenablement, les écarts relatifs entre la valeur réelle du total d'après la procédure officielle $T(X_j)$ et l'estimation du modèle $\hat{T}(X_j)$ sous la forme $B(X_j) = (\hat{T}(X_j) - T(X_j))/T(X_j)$, pour $j = 1, 2$, sont $B(X_1) = 0,005$ et $B(X_2) = 0,002$. Ces valeurs ne sont pas directement comparables au seuil de tolérance de $\delta = 0,001$; en fait, ce seuil a trait uniquement à l'effet des erreurs d'unité de mesure encore présentes, tandis que $B(X_j)$ est également affecté par d'autres formes d'erreurs. Donc, pour une

estimations et causer un biais important. Afin de sélectionner les unités éventuellement erronées qui sont les plus susceptibles d'avoir un effet important sur le *vérification sélective*. Soit X_1, X_2 les variables TS et TI, respectivement. Pour chaque unité $u_i, i = 1, \dots, n$, et pour chaque variable $X_j, j = 1, 2$, définissons :

X_{ij} : données dépourvues d'erreur systématique;
 Y_{ij} : données observées;

\tilde{X}_{ij} : données après le traitement de l'erreur systématique d'après la classification au moyen d'un modèle de mélanges (c'est-à-dire $\tilde{X}_{ij} = Y_{ij}$ ou $\tilde{X}_{ij} = Y_{ij}/1\,000$ selon la grappe à laquelle l'unité u_i est assignée).

Supposons que les estimations cibles soient les totaux de population $T(X_j) = \sum_i X_{ij}$. En outre, représentons par $E_{ij}(\cdot)$ l'espérance sur la distribution de la variable aléatoire X_{ij} , sachant les données observées X_{ij} et les données après correction \tilde{X}_{ij} . Alors, il découle de l'inégalité

$|\sum_i E_{ij}(X_{ij} - \tilde{X}_{ij})| \leq \sum_i E_{ij}|X_{ij} - \tilde{X}_{ij}|$ que la quantité dans le deuxième membre peut être considérée comme une borne supérieure du biais probable de l'estimation du total pour la variable X_j fondée sur les valeurs corrigées \tilde{X}_{ij} .

La dernière considération donne à penser à une méthode pour sélectionner les unités « influentes » en ce qui concerne l'estimation $T(X_j)$: afin de garantir le niveau requis d'exactitude et de réduire au minimum les coûts de la

vérification manuelle, nous définissons une fonction de score local $S_j^i = (E_{ij}|X_{ij} - \tilde{X}_{ij}| / \tilde{T}(X_j))$, où $\tilde{T}(X_j)$ est une estimation de référence pour $T(X_j)$, par exemple l'estimation provenant d'une enquête antérieure, ou une estimation robuste. Dans notre cas, afin de rendre robuste l'estimation préliminaire, nous commençons par exclure des données les observations atypiques, puis nous calculons la valeur moyenne sur ce sous-ensemble, et nous la multi-

plions par le nombre total d'unités. Le score local S_j^i mesure l'effet de l'erreur d'unité de mesure éventuellement associée à l'unité u_i sur l'estimation cible $T(X_j)$. Alors, nous pouvons trier les unités en fonction de leur score S_j^i et, en commençant par les valeurs les plus élevées, sélectionner les premières unités jusqu'à ce que la somme des valeurs S_j^i restantes soient inférieures à un seuil préalable.

Si nous considérons simultanément les deux variables TS et TI, nous pouvons obtenir un score global S_i^j pour $i = 1, \dots, n$, en combinant comme il convient les fonctions de score local S_{ij}^j , $j = 1, 2$. Les choix possibles sont $S_i^j = (S_{i1}^1 + S_{i2}^2)/2$, ou $S_i^j = \max_{j=1,2} S_{ij}^j$. Par exemple, la dernière fonction assure que l'effet de l'erreur d'unité de

ne soit pas supérieur à S_i^j . Afin de calculer les scores S_{ij}^j , nous devons estimer l'espérance conditionnelle $E_{ij}|X_{ij} - \tilde{X}_{ij}|$ pour chaque unité $u_i, i = 1, \dots, n$, et pour chaque variable $X_j, j = 1, 2$, ce qui peut se faire facilement au moyen des probabilités a posteriori. Par exemple, supposons que l'unité u_i ait été assignée à la grappe G_2 . Cela signifie que, pour cette unité, la valeur observée de TS (Y_{i1}^1) a été considérée correcte, tandis que la valeur observée de TI (X_{i2}^2) a été considérée comme étant affectée d'une erreur d'unité de mesure (c'est-à-dire multipliée par 1 000). La correction consiste à diviser par 1 000 la valeur observée de TI, c'est-à-dire ($X_{i1}^1 = Y_{i1}^1, \tilde{X}_{i2}^2 = X_{i2}^2/1\,000$). L'espérance conditionnelle $E_{ij}|X_{ij} - \tilde{X}_{ij}|$ peut être calculée comme suit :

$$E_{ij}\left[X_{ij} - \tilde{X}_{ij}\right] = |Y_{i1}^1 - Y_{i1}^1| \Pr(u_i \in G_1 \cup G_2) + \left|\frac{Y_{i1}^1}{1\,000} - Y_{i1}^1\right| \Pr(u_i \in G_3 \cup G_4) = \frac{1}{999} Y_{i1}^1 (\hat{\tau}_{3i} + \hat{\tau}_{4i}) \\ E_{ij}\left[X_{i2}^2|X_{i2}^2 - \tilde{X}_{i2}^2\right] = \left|\frac{X_{i2}^2}{1\,000} - \frac{X_{i2}^2}{1\,000}\right| \Pr(u_i \in G_2 \cup G_4) + \left|X_{i2}^2 - \frac{X_{i2}^2}{1\,000}\right| \Pr(u_i \in G_1 \cup G_3) = \frac{1}{999} X_{i2}^2 (\hat{\tau}_{1i} + \hat{\tau}_{3i}),$$

où $\hat{\tau}_k$ est la probabilité estimée que l'unité u_i appartienne à la grappe G_k . De façon semblable, nous pouvons calculer les fonctions de score pour toutes les unités. En pratique, dans notre application, nous trions les unités en fonction de leur score global $S_i^j = \max_{j=1,2} S_{ij}^j$ (ordre ascendant). Puis, nous excluons de l'examen manuel toutes les premières observations, de sorte que la somme cumulée de leurs S_i^j soit inférieure à δ , où δ est un seuil de tolérance spécifié pour l'effet sur les estimations des erreurs encore présentes dans les données. À la figure 6, nous présentons le comportement de la somme cumulée de $S_i^j, S_i^{(i)} = \sum_{j=1,2} S_{ij}^j$, pour les 10 premières observations de S_i^j , pour la plupart d'entre elles, $S_i^{(i)}$ est proche de zéro, parce que produit une image illisible en ce qui concerne les différences de grandeur. Notons que nous prévoyons une erreur relative résiduelle inférieure à $\delta = 0,001$ en sélectionnant uniquement les deux premières unités (représentées par des croix).

Tableau 3 Nombres d'observations critiques (OC) et d'unités mal classées (MC) pour trois seuils distincts de probabilité de classification

β	MN-Mixt			MT-Mixt			ST-Mixt		
	Pr. Class – OC	Pr. Class – MC	Pr. Class – OC	Pr. Class – MC	Pr. Class – OC	Pr. Class – MC	Pr. Class – OC	Pr. Class – MC	Pr. Class – MC
0,99	119	19	63	12	182	26			
0,975	76	18	46	11	82	26			
0,95	55	14	35	9	66	21			

Tableau 4 Nombres d'observations critiques (OC) et d'unités mal classées (MC) pour l'atypicité et la probabilité de classification

Seuils	MN-Mixt		MT-Mixt		ST-Mixt	
	OC	MC	OC	MC	OC	MC
$\alpha = 0,005, \beta = 0,975$	84	18	79	14	98	24

4.2 Application à des données réelles : Le Système d'enquêtes sur l'eau de l'Italie de 1999

À la présente section, nous décrivons une application de l'approche du modèle de mélanges à des données d'enquête réelles. Ces données sont tirées du Système d'enquêtes sur l'eau (SEE) de l'Italie de 1999. Le SEE est un recensement visant à recueillir des renseignements sur le prélèvement, la fourniture et la consommation d'eau dans les 8 100 municipalités italiennes. Nous limitons l'analyse aux municipalités appartenant à l'un des domaines de données définis par altimétrie (2 041 observations) et aux variables principales *Total de l'eau facturée* (TT) et *Total de l'eau fournie* (TS). Ces variables ont trait toutes deux à des volumes d'eau et il est demandé aux répondants de déclarer ces volumes en milliers de mètres cubes. Le diagramme de dispersion en échelle logarithmique de la quantité d'eau facturée par habitant (WT) en fonction de la quantité d'eau fournie par habitant (WT) (figure 4) montre l'existence de quatre groupes correspondant à une erreur d'unité de mesure dans l'une, les deux ou ni l'une ni l'autre des deux variables cibles. Ces erreurs sont probablement dues à un malentendu chez certains répondants qui ont exprimé les volume en litres ou en mètres cubes plutôt qu'en milliers de mètres cubes, comme il l'était demandé. Comme il fallait s'y attendre, les deux groupes dont la population est la plus nombreuse sont celles correspondant aux unités non erronées et aux unités où les deux variables sont erronées. Néanmoins, nous observons la présence de deux groupes rares correspondant aux observations où l'erreur d'unité de mesure a trait uniquement à TT ou à TS, respectivement.

Dans le tableau 5, une étiquette est attribuée à chaque groupe associée à un patron d'erreur particulier. Par souci de simplicité, nous introduisons deux drapeaux E_{TS} et E_{TT} dont la valeur est égale à 1 ou 0, selon que la variable

correspondante est affectée ou non par l'erreur d'unité de mesure ou non.

Afin de dépitier et de corriger les erreurs d'unité de mesure, nous appliquons la procédure décrite aux sections 2 et 3. Nous classons chaque observation en fonction d'un patron d'erreur particulier, autrement dit nous affectons chaque unité à l'une des grappes G_i , pour $i = 1, \dots, 4$. Les résultats sont présentés au tableau 6.

Pour chaque unité, nous calculons aussi l'indice d'atypicité et nous choisissons le seuil $\alpha = 0,005$ afin de repérer les unités atypiques. Pour ce seuil, 71 observations sont sélectionnées comme étant atypiques, et marquées par des « croix » dans la figure 7. Après avoir calculé les valeurs \hat{a}_{8i} conformément à la formule (3), nous pouvons faire un test pour évaluer l'hypothèse de normalité. En fait, à l'instar de McLachlan et Basford (1988, chapitre 2), nous appliquons le test d'Anderson-Darling de l'uniformité de \hat{a}_{8i} à chaque grappe estimée individuelle. La valeur p est inférieure à 0,001 pour les deux plus grandes grappes. Puisque le test est fondé sur des approximations asymptotiques, nous ne tenons pas compte des résultats obtenus pour les deux autres populations rares. À la figure 5, nous donnons les quantiles empiriques d'échantillon en fonction des quantiles normaux des variables $\log(WT)$ et $\log(WS)$, en nous concentrant uniquement sur le sous-ensemble de données classées comme étant non erronées. Nous constatons que l'écart par rapport à la loi normale est du principalement à la section 4.1, où la méthode a donné des résultats satisfaisants également dans des conditions non gaussiennes, nous nous attendons à ce que l'approche du mélange de lois donne de bons résultats pour les données d'enquête. Les résultats de l'application illustrés ci-après confirment ce comportement.

En ce qui concerne l'atypicité, nous constatons que, si le modèle est spécifié correctement, le rôle de l'indice d'atypicité dans la correction des unités mal classées est négligeable, tandis que les probabilités de classification donnent de meilleurs résultats. Par ailleurs, l'indice d'atypicité est important si le modèle dévie de la normalité. Il est clair que le nombre d'observations sélectionnées pour une combinaison de seuils α et β n'est pas égal à la somme des fréquences obtenues dans les tableaux 2 et 3. Donc, afin d'évaluer l'effet collectif des deux indices, nous choisissons les deux seuils suivants $\alpha = 0,005$ et $\beta = 0,975$. Nous présentons à la figure 3 (deuxième ligne) les unités sélectionnées pour la valeur d'atypicité seulement (carrés), et pour la probabilité de classification seulement (triangles) et pour les deux (croix) dans MN-Mixt(a), MT-Mixt(a), ST-Mixt(a) et d'unités mal classées (MC) pour trois seuils distincts d'atypicité

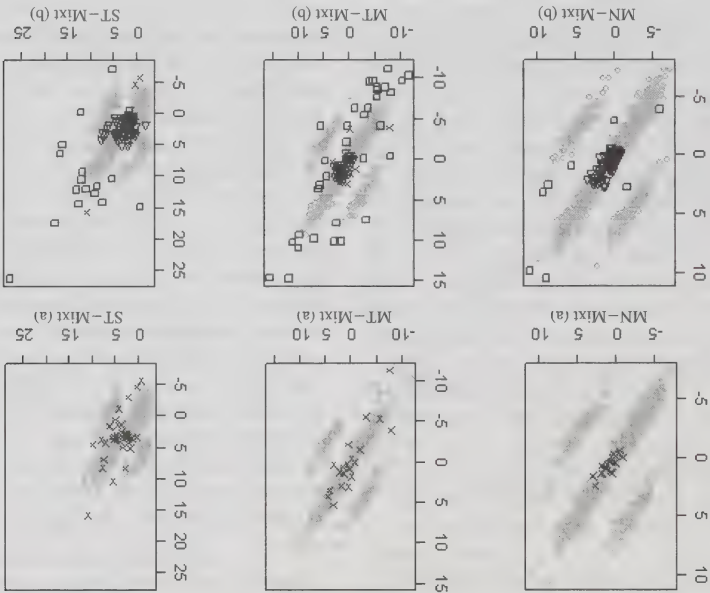


Figure 3. Unités mal classées (croix) dans MN-Mixt(a), MT-Mixt(a) et ST-Mixt(a). Unités critiques pour l'atypicité (carrés), pour la probabilité de classification (triangles) et pour les deux (croix) dans MN-Mixt(b), MT-Mixt(b) et ST-Mixt(b).

Tableau 2

Nombres d'observations critiques (OC) et d'unités mal classées (MC) pour trois seuils distincts d'atypicité

α	MN-Mixt Aryp - OC	MT-Mixt Aryp - MC	ST-Mixt Aryp - MC
0,05	50	84	68
0,01	15	50	33
0,005	8	39	20
0,001	4	25	14

ST-Mixt), correspondant aux trois populations distinctes MN, MT et ST, respectivement.

Pour chaque échantillon, nous calculons le nombre de classifications correctes obtenues en utilisant le modèle de mélanges décrit à la section 2. Le nombre moyen de classifications correctes sur les 100 échantillons est présenté au tableau 1.

L'examen du tableau 1 montre que la fréquence des classifications correctes diminue lorsque la déviation par rapport à la loi normale augmente. Cependant, elle semble acceptable même dans le cas critique ST où la population est caractérisée par une distribution à la fois asymétrique et à queues lourdes.

Tableau 1

Fréquence des classifications correctes			
<hr/>			
MN	MT	ST	
<hr/>			
% correctement classée			
98,5	97,5	95,6	

Comme nous en avons discuté à la section 3, l'approche du modèle de mélanges fournit des éléments (tels que le degré d'atypicité et la probabilité de classification) qui

peuvent être utilisés pour déterminer l'ordre de priorité des unités pour l'examen manuel. Par conséquent, une évaluation globale de la procédure devrait tenir compte des résultats d'une approche de vérification sélective fondée sur ces diagnostics du modèle.

Afin d'analyser les caractéristiques de l'indice d'atypicité et de la probabilité de classification, nous examinons un seul échantillon de 1 000 observations tiré à partir des trois populations présentes jusqu'à présent. La figure 3 illustre les trois échantillons MN-Mixt(a), MT-Mixt(a) et ST-Mixt(a), où les unités classées incorrectement sont représentées par une croix sur le même graphique. Le nombre d'unités classées incorrectement est 19 pour MN-Mixt, 20 pour MT-Mixt et 36 pour ST-Mixt.

Pour cet échantillon, nous nous concentrons sur l'effet de divers seuils pour l'atypicité (α) et la probabilité de classification (β). Pour chaque seuil, nous présentons aux tableaux 2 et 3 le nombre d'unités situées sous le seuil, c'est-à-dire le nombre d'observations critiques (Atyp. - OC, Pr. class. - OC), et parmi ces observations, le nombre d'unités mal classées (Atyp. - MC, Pr. class. - MC).

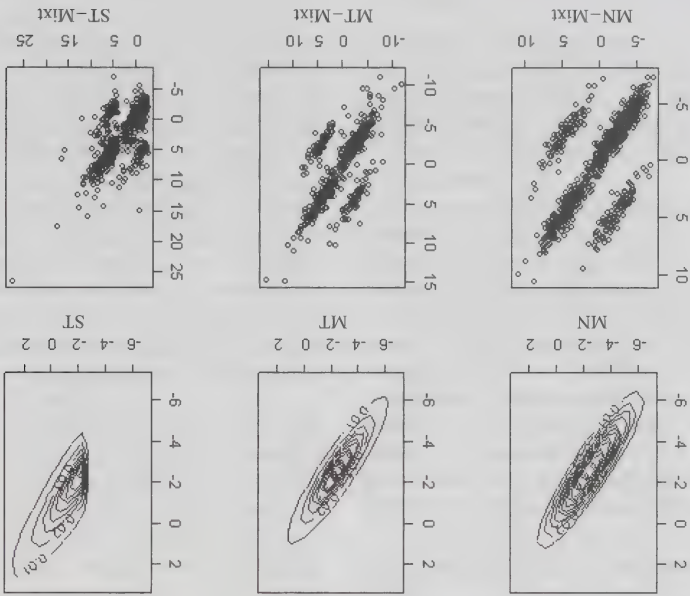


Figure 2. Tracés de contours des trois distributions bivariées : multinomiale (MN), t de Student (MT), t asymétrique (ST), et diagrammes de dispersion des mélanges correspondant MN-Mixt, MT-Mixt et ST-Mixt.

\hat{a}^{gt} comme étant l'aire à la droite de la valeur \hat{p}^{gt} sous la distribution $F^{q,n}$ (pour des précisions, consulter McLachlan et Basford 1988, chapitre 2).

Sous l'hypothèse de normalité, \hat{a}^{gt} pour $t = 1, \dots, m_g$ est approximativement uniformément distribué sur l'intervalle (0,1). Hawkins (1981) propose d'utiliser la statistique d'Anderson-Darling pour évaluer la distribution uniforme de \hat{a}^{gt} . Les \hat{a}^{gt} sont également utiles pour détecter les valeurs extrêmes, c'est-à-dire les observations aberrantes par rapport au modèle. Dans McLachlan et Basford (1988), la probabilité que y^{gt} soit atypique est d'autant plus élevée que \hat{a}^{gt} est faible, si bien que toutes les observations avec $\hat{a}^{gt} < \alpha$, où α est un seuil spécifié, peuvent être considérées comme étant atypiques. Les valeurs proposées du seuil varient de $\alpha = 0,05$ à $\alpha = 0,005$, selon les observations aberrantes (valeurs plus ou moins extrêmes) qu'il faut sélectionner.

3. Diagnostics pour la vérification sélective

Une fois que les paramètres du mélange sont estimés, nous pouvons classer les données dans les diverses grappes; autrement dit, pour chaque observation, nous pouvons déterminer s'il s'agit d'une erreur ou non, et sur quelle variable l'erreur porte. Cependant, divers types d'observations critiques peuvent être définis après la phase de modélisation, à savoir les unités classées dans une grappe, mais ayant une probabilité non négligeable d'appartenir à une autre grappe, et les observations qui sont des valeurs aberrantes par rapport au modèle.

Afin d'accroître l'exactitude des données, il serait utile de procéder à une double vérification des observations critiques (par examen manuel, ou, dans les cas les plus d'efficients, par un suivi). Par ailleurs, afin de réduire la survérification éventuelle et les coûts de vérification, il convient de concentrer l'examen manuel (et/ou) le suivi sur les observations les plus critiques. Le modèle de mélanges proposé fournit des diagnostics directs que l'on peut utiliser à cette fin. Un premier type d'unités critiques est représenté par les observations éventuellement classées incorrectement. Afin de déterminer le degré de confiance dans la classe attribuée à une observation y_i^t , nous pouvons considérer la probabilité correspondante résultant de (2). Les observations pour lesquelles cette probabilité n'est pas très proche de l'unité ont une probabilité non négligeable d'appartenir à une autre grappe. Ces observations sont celles situées dans la région où les composantes du mélange se superposent. En plus du type susmentionné d'unités critiques, il existe d'autres observations qui sont éloignées de toutes les grappes (toutes les composantes du mélange), c'est-à-dire les valeurs aberrantes par rapport au modèle. Ces observations représentent aussi des situations critiques. Afin

de repérer ce genre de valeur aberrante, nous nous servons des quantités \hat{a}_{ij}^{gt} décrites à la section précédente. La probabilité de classification et l'indice d'atypicalité \hat{a}^{gt} devraient être utilisés, conformément à une approche de vérification sélective/section l'importance (Laouche et Berthelot 1992; Lawrence et McKenzie 2000), pour conclure des fonctions de score appropriées afin de déterminer l'ordre de priorité des unités critiques. Nous donnons un exemple d'utilisation de ces diagnostics dans ce but à la sous-section 4.2.

4. Exemples

Nous décrivons ici certaines expériences réalisées en vue d'étudier les particularités de la méthode proposée. En premier lieu, grâce à une étude en simulation, nous analysons les propriétés du modèle proposé lorsqu'il est appliqué à des données qui s'écartent de la normalité. Deuxièmement, au moyen de données réelles, nous décrivons comment l'approche peut être appliquée dans le domaine de la statistique officielle.

Toutes les expériences sont réalisées dans l'environnement R pour calcul statistique (<http://www.r-project.org/>).

4.1 Exemple simulé : Déviation par rapport à la loi normale

Dans cette expérience, nous décrivons les résultats obtenus en appliquant la méthode du mélange de lois aux trois populations distinctes illustrées à la première ligne de la figure 2. La première distribution est une loi normale bivariée (MN), qui représente donc le cas où le modèle est spécifié correctement. La deuxième correspond à une loi t bivariée (MT), c'est-à-dire qu'elle mime la situation où la déviation par rapport à la loi normale se résume essentiellement à des queues plus lourdes. La dernière est une loi t asymétrique bivariée (ST) (Azzalini et Capitanio 2003; Azzalini, Dal Cappello et Kozi 2003), qui représente une population distribuée conformément à une loi asymétrique à queues lourdes. Pour ces distributions, nous construisons un modèle de mélanges à quatre composantes en ajoutant à chaque unité l'un des quatre vecteurs de translation $C_1 = (0, 0)$, $C_2 = (0, \log(1\,000))$, $C_3 = (\log(1\,000), 0)$ ou $C_4 = (\log(1\,000), \log(1\,000))$, avec les probabilités $\pi_1 = 0,5$, $\pi_2 = 0,1$, $\pi_3 = 0,1$ et $\pi_4 = 0,3$, respectivement. Ces paramètres représentent les proportions de mélange du modèle et ont trait, respectivement, aux probabilités de l'absence de translation dans les variables, d'une translation dans une des deux variables seulement et d'une translation dans les deux variables, respectivement. À partir de chaque observation, nous tirons 100 échantillons de 1 000 observations. À la deuxième ligne de la figure 2 nous présentons l'un de ces échantillons (MN–Mixt, MT–Mixt et

$$\tau_i(y_i; \theta, \pi) < \tau_g(y_i; \theta, \pi) \quad g = 1, \dots, h; g \neq i.$$

La règle d'affectation qui précède est la solution optimale du problème de classification, en ce sens qu'elle minimise le taux global d'erreur (Anderson 1984, chapitre 6). Puisque, au lieu des paramètres (θ, π) , généralement inconnus, nous utilisons les estimations du maximum de vraisemblance $(\hat{\theta}, \hat{\pi})$, la règle de classification devient :

$$\tau_i(y_i; \hat{\theta}, \hat{\pi}) < \tau_g(y_i; \hat{\theta}, \hat{\pi}) \quad g = 1, \dots, h; g \neq i. \quad (2)$$

Nous supposons que la fonction $f_i^g(y_i; \theta)$ est une fonction de densité multivariée $MN(\mu^g, \Sigma)$ et que chaque fonction $\phi_g(\cdot)$ agit sur le vecteur de moyennes μ comme une translation : $\phi_g(\mu) = \mu + C^g$, où C^g représente le vecteur de translation pour la moyenne de la g^e grappe que nous supposons être connue. Ce cadre, comme nous l'avons déjà souligné, convient pour le traitement de l'erreur d'unité de mesure. Afin de calculer les estimations de vraisemblance, nous utilisons l'algorithme EM, tel que proposé dans McLachlan et Basford (1988). Néanmoins, un effort supplémentaire est nécessaire pour adapter l'algorithme à notre situation particulière, où les vecteurs moyens des composantes du mélange sont liés par une relation fonctionnelle connue. Donc, alors que dans le cas sans contrainte (McLachlan et Basford 1988), un vecteur de moyennes distinct doit être estimé pour chaque composante du mélange, dans notre situation contraire, il suffit d'en estimer un seul. L'algorithme EM modifié résultant consiste à définir une valeur initiale estimée au jugé pour les paramètres à estimer $\hat{\pi}_{(0)}^g$ pour $g = 1, \dots, h$, $\hat{\pi}_{(0)}$ et à appliquer jusqu'à la convergence le schéma récursif suivant :

- i) calculer les probabilités a posteriori $\tau_{gi}^{(k)} = \tau_{(k)}^g(y_i; \theta, \pi)$ sous les estimations courantes $\hat{\pi}_{(k)}^g$, $\sum_{g=1}^h \hat{\pi}_{(k)}^g$ (k est l'indice supérieur désignant le k^e cycle)

$$\hat{\pi}_{gi}^{(k)} = \frac{\hat{\pi}_{(k)}^g \exp\left\{-\frac{1}{2}\left(y_i - \hat{\mu}_{(k)}^g\right)' \left(\sum_{(k)}^{-1}\right) \left(y_i - \hat{\mu}_{(k)}^g\right)\right\}}{\sum_{g=1}^h \hat{\pi}_{(k)}^g \exp\left\{-\frac{1}{2}\left(y_i - \hat{\mu}_{(k)}^g\right)' \left(\sum_{(k)}^{-1}\right) \left(y_i - \hat{\mu}_{(k)}^g\right)\right\}}$$

- ii) calculer les nouvelles estimations au moyen des équations récursives suivantes :

$$\hat{\pi}_{(k+1)}^g = \sum_{i=1}^n \hat{\tau}_{gi}^{(k)} / n$$

$$\hat{\mu}_{(k+1)}^g = \sum_{i=1}^n \sum_{g=1}^h \hat{\tau}_{gi}^{(k)} y_i / n - \sum_{g=1}^h C^g \hat{\pi}_{gi}^{(k+1)}$$

Souignons que $\hat{\mu}_{(k)}^g$ représente $\mu_{(k)}^g + C^g$.

Dans les applications pratiques, le choix des points de départ s'avère essentiel, comme d'habitude dans les algorithmes EM (voir Biernacki, Celeux et Govaert 2003). Pour surmonter ce problème, nous utilisons une stratégie d'initialisation, inspirée de Biernacki et coll. (2003), qui consiste en plusieurs exécutions brèves, en ce qui concerne le nombre d'itérations, de l'algorithme provenant des initialisations aléatoires, suivies par une longue exécution de l'algorithme EM provenant de la solution qui maximise la log-vraisemblance observée.

Il mérite d'être mentionné qu'à cause des contraintes d'emplacements, les paramètres qui doivent être estimés sont sensiblement moins nombreux que ceux d'un problème de mélange de lois habituel. En fait, cette différence est d'autant plus importante que le nombre de variables à analyser est grand; par exemple, dans le cas de 3 variables et de 8 grappes, nous devons estimer 16 paramètres au lieu de 37. Cet aspect est particulièrement important quand nous avons affaire à de petits échantillons. De surcroît, les contraintes sur les emplacements des grappes permettent de repérer plus facilement les « grappes rares ». En fait, les distances relatives entre les vecteurs de moyennes étant fixes, le problème d'estimation se réduit à estimer l'emplacements des grappes. Autrement dit, puisque l'emplacement d'un centroïde détermine sans ambiguïté les positions de tous les autres, les paramètres des petites grappes sont estimés plus facilement que s'il n'existait pas de contrainte.

Puisque la modélisation présentée est fondée sur l'hypothèse que les observations suivent une loi normale, la validation du modèle est une question dont il faut tenir compte. Le problème de l'évaluation de la normalité dans les modèles de mélanges de lois est bien décrit dans McLachlan et Basford (1988). Il est essentiellement fondé sur les quantités $\hat{a}_{gi}^{(k)}$ décrites plus bas. Soit y_{gi} pour $i = 1, \dots, m_g$, les observations assignées à la g^e grappe pour $g = 1, \dots, h$, conformément au modèle estimé. Soit \hat{p}_{gi} la valeur calculée en se servant des paramètres estimés, au moyen de la formule :

$$\hat{p}_{gi} = \frac{(v m_g / q) D \left(y_{gi}; \hat{\mu}_{(k)}^g \right) \left(\sum_{(k)}^{-1} \right) \left(y_{gi} - \hat{\mu}_{(k)}^g \right) \left(\sum_{(k)}^{-1} \right) \left(y_{gi} - \hat{\mu}_{(k)}^g \right)}{(v + q) (m_g - 1) - m_g D \left(y_{gi}; \hat{\mu}_{(k)}^g \right)}$$

où $D(\cdot, \cdot; M)$ est le carré de la distance de Mahalanobis basée sur la mesure M , et $v = n - h - q$. Nous définissons

données réelles. Enfin, à la section 5, nous présentons nos conclusions et les futures travaux de recherche.

2. Le modèle

Il est difficile de donner une formalisation complète des erreurs aléatoires et systématiques. Dans le présent contexte, nous fournissons une définition qui, bien qu'elle ne soit pas exhaustive, inclut un grand nombre de situations courantes. Soit X^* le vecteur de variables cibles de l'enquête et (μ, Σ^*) le vecteur de moyennes correspondant et la matrice des covariances. Supposons que le processus de mesure soit affecté par un mécanisme d'erreur aléatoire R ayant un effet sur la structure de covariance de X^* , mais laissant le vecteur de moyennes inchangé et, conséquemment, représentations par X la variable « combinée » correspondant à son espérance mathématique $\mu = \frac{1}{S} \phi(\mu)$ pour une certaine fonction ϕ (par exemple, si l'on suppose que le mécanisme d'erreur est additif, $\phi(\mu) = \mu + \text{constante}$). En raison des deux mécanismes d'erreur, que nous supposons être indépendants l'un de l'autre, nous pouvons écrire les données observées au moyen d'un vecteur aléatoire X dont le mécanisme d'erreur systématique. Notre façon d'aborder le traitement des erreurs systématiques consiste à construire pour X un modèle axé uniquement sur la détection des erreurs systématiques, donc visant à récupérer les données aléatoirement contenues représentées par le vecteur aléatoire X . Cette approche est celle généralement adoptée dans les procédures de vérification où les erreurs systématiques et les erreurs aléatoires sont traitées séparément et hiérarchiquement.

La définition donnée plus haut de l'erreur systématique comprend l'erreur de mesure, une fois que les données ont été transformées en échelle logarithmique. En fait, l'erreur d'unité de mesure a habituellement l'effet de multiplier les variables par un facteur constant. Donc, sur une échelle logarithmique, les données erronées apparaissent comme la translation d'un vecteur de constantes qui dépend des items qui sont erronés (« patron d'erreur »), tandis que la structure de covariance est la même pour chaque patron d'erreur. Qui plus est, en fait, les variables des enquêtes-entreprises sont fréquemment considérées comme suivant une loi log-normale. Donc, en échelle logarithmique, nous pouvons adopter les paramètres gaussiens.

Partant de la formalisation exposée jusqu'à présent, notre but est maintenant d'affecter chaque observation à un « patron d'erreur » particulier, ce qui revient à localiser les items entachés d'une erreur. Si nous interprétons chaque

patron d'erreur comme étant une « grappe », le problème de localisation de l'erreur devient un problème de classification (cluster analysis) et nous pouvons profiter des enseignements de la théorie de la classification basée sur un modèle

(Frailay et Raftery 2002).

Plus précisément, supposons que nous ayons n observations indépendantes $X_i = (X_{i1}, \dots, X_{iq})$, $i = 1, \dots, n$, correspondant aux vecteurs de dimension q représentés par $X_i = (X_{i1}, \dots, X_{iq})$ avec la f.d.p. $f(x_1, \dots, x_q; \theta)$, telle que $E(X_i) = (\mu_1, \dots, \mu_q)$ et $\text{Var}(X_i) = \Sigma$.

Si nous supposons que le seul effet des erreurs systématiques sur le vecteur aléatoire X est la transformation de son espérance μ en $\phi(\mu)$, où $\phi^g(\cdot): R^g \rightarrow R^g$, pour $g = 1, \dots, h$, est un ensemble de fonctions connues, les fonctions ϕ_g caractérisent de façon univoque h grappes (patrons d'erreur) distinctes, qui ne diffèrent l'une de l'autre que par le paramètre d'emplacement. Par exemple, si l'erreur systématique affectait toutes les variables X_s , pour $s = 1, \dots, q$, de la même façon en transformant leur espérance μ_s conformément à $\mu_s \rightarrow \mu_s + C$, où C est une constante connue, le nombre de grappes serait $h = 2^q$, c'est-à-dire le nombre de combinaisons différentes de l'occurrence de l'erreur sur les q variables (y compris le cas où il n'y a pas d'erreur). Dans ce cas, chaque fonction ϕ_g et chaque grappe correspondante sera associée à l'un des 2^q sous-ensembles possibles de variables affectées par l'erreur; par exemple, le groupe G caractérisé par le vecteur de moyennes $\mu_G = (\mu_1, \mu_2 + C, \mu_3, \mu_4, \dots, \mu_q)$ est une grappe d'unités présentant une erreur n'affectant que la variable X_2 . Soulignons que nous supposons qu'il existe un matrice des covariances commune, parce que nous émettons l'hypothèse que l'erreur aléatoire possible agit de la même façon sur toutes les données.

Aux fins de la localisation de l'erreur, nous suivons une approche par modèle basée sur des modèles de mélanges finis de lois où chaque composante du mélange G_g , $g = 1, \dots, h$, représente un patron d'erreur particulier. Formellement, nous supposons que les $X_i = (X_{i1}, \dots, X_{iq})$, pour $i = 1, \dots, n$, sont i.i.d. par rapport à $\sum_{i=1}^n \pi_i f_i(\cdot; \theta_i)$, où $\sum_i \pi_i = 1$ et $\pi_i \geq 0$. Les paramètres de mélange π_i représentent la probabilité qu'une observation appartienne à la i^{e} composante du mélange.

Afin de classer une observation y_i dans l'un des h groupes, nous calculons la probabilité a posteriori $\tau_g(y_i; \theta, \pi) = \text{pr}(i^{\text{e}} \text{ observation} \in G_g | y_i; \theta, \pi)$, c'est-à-dire

$$\tau_g(y_i; \theta, \pi) = \pi_g f_g(y_i; \theta_g) / \sum_{i=1}^h \pi_i f_i(y_i; \theta_i) \quad (1)$$

La i^{e} observation est assignée à la grappe G_i

La CMO est une enquête par sondage périodique conçue pour recueillir des renseignements sur l'emploi, le nombre d'heures travaillées, les traitements et salaires, et le coût de la main-d'œuvre auprès d'environ 12 000 entreprises composées de dix employés. La figure 1 représente le logarithme du coût de la main-d'œuvre (LCOST), du nombre d'emplois (LEMPLOY) et du nombre d'heures travaillées (LWORKEDH) dans une matrice de diagrammes de dispersion. Notons qu'à l'étape de la vérification, la variable d'emploi est sans erreur, en raison d'une vérification préliminaire de l'information provenant des registres des entreprises (Cittanni, Di Zio, Luzi et Seeber 2000). L'analyse de la figure 1 montre que le coût de la main-d'œuvre est affecté par deux types d'erreur d'unité de mesure (c'est-à-dire facteurs de 1 million et de 1 000), tandis que le nombre d'heures travaillées ne présente que l'erreur du facteur 1 000. Ces erreurs donnent lieu à la formation de diverses grappes dans la figure 1. Il convient de souligner, que, dans chaque diagramme de dispersion, les grappes qui se trouvent dans le coin inférieur gauche représentent les données non erronées.

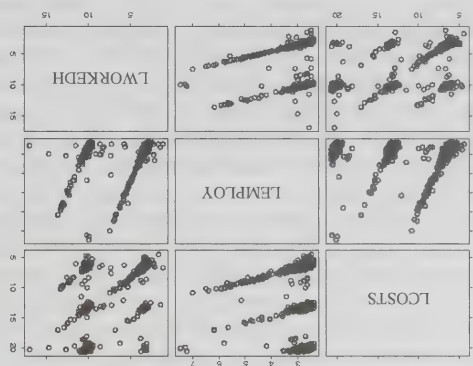


Figure 1. Diagramme de dispersion multiple du coût total de la main-d'œuvre, du nombre d'emplois et du nombre d'heures travaillées (échelle logarithmique).

L'exemple du SBE sera décrit en détail à la sous-section 4.2, où nous présenterons une application de la méthode proposée dans le présent article en vue de repérer et de traiter les erreurs d'unité de mesure.

Dans le cas de l'erreur d'unité de mesure, l'élément essentiel est la localisation des items erronés plutôt que leur traitement. En fait, une fois qu'un item est classé comme étant erroné, le traitement optimal est déterminé de façon unique et consiste à prendre une mesure déterministe de correction de la valeur originale par une opération inverse

(par exemple, division par 1 000) qui neutralise l'effet de l'erreur.

En général, on s'attaque à l'erreur d'unité de mesure par des procédures ponctuelles s'appuyant essentiellement sur la représentation graphique de distributions marginales ou bivariées, et sur des *vérifications de rapport*. Une vérification de rapport est une règle énonçant que la valeur d'un rapport donne entre deux variables doit être comprise dans un intervalle prédéfini. Les bornes de l'intervalle sont généralement déterminées d'après des renseignements a priori ou grâce à une analyse exploratoire des données, en utilisant éventuellement des données auxiliaires fiables. Pour le genre d'erreur susmentionné, les vérifications de rapport sont efficaces si les données sur l'une des deux variables ne contiennent pas d'erreur. En outre, elles ne permettent de tenir compte que des relations bivariées entre variables et, même si l'on recourt à l'inspection graphique interactive (par exemple, matrice de diagrammes de dispersion), on doit se limiter à une analyse par paire, et ne pas tenir compte des interactions plus complexes entre les variables. Enfin, il faut souligner que le recours à l'analyse par paire implique que les variables doivent être traitées selon une hiérarchie prédéterminée, ce qui accroît la complexité de la méthode de localisation de l'erreur.

Si l'on s'en tient aux approches classiques, le problème de localisation de l'erreur est non seulement complexe, mais également long et coûteux. La durée et le coût dépendent principalement : 1) de la complexité de la conception et de la mise en œuvre de procédures déterministes automatisées *ponctuelles* et 2) des ressources consacrées à la vérification manuelle des observations dont la probabilité d'être erronées est faible (et/ou) dont l'effet sur les estimations cibles est faible (*survérification*).

Dans le présent article, nous proposons une formalisation probabiliste du problème au moyen de modèles de mélanges finis (McLachlan et Basford 1988; McLachlan et Peel 2000). Cette modélisation peut offrir une approche statistique disciplinée, permettant d'estimer la probabilité conditionnelle qu'une observation soit affectée par une erreur d'unité de mesure. L'avantage de l'approche proposée est qu'elle représente une méthode générale permettant de faire une analyse multivariée des données et fournissant des éléments qui peuvent être utilisés pour optimiser l'équilibre entre les composantes automatiques et interactives de la procédure de vérification, c'est-à-dire entre la durée et l'exactitude (Granquist et Kovar 1997).

La présentation de l'article est la suivante. À la section 2, nous décrivons le modèle proposé, ainsi que l'algorithme EM utilisé pour estimer les paramètres. À la section 3, nous décrivons les diagnostics pour la vérification sélective. À la section 4, nous illustrons les résultats de l'application de la méthode proposée à des données simulées, ainsi que des

Vérification des erreurs systématiques d'unité de mesure au moyen de la modélisation par mélanges

Marco Di Zio, Ugo Guarnera et Oretta Luzi¹

Résumé

Dans le domaine de la statistique officielle, le processus de vérification des données joue un rôle important dans la rapidité de production, l'exactitude des données et les coûts d'enquête. Les techniques adoptées pour détecter et éliminer les erreurs que contiennent les données doivent essentiellement tenir compte simultanément de tous ces aspects. L'une des erreurs systématiques que l'on observe fréquemment dans les enquêtes visant à recueillir des données numériques est celle de l'unité de mesure. Cette erreur a une forte incidence sur la rapidité de production, l'exactitude des données et le coût de la phase de vérification et d'imputation. Dans le présent article, nous proposons une formalisation probabiliste du problème basée sur des modèles de mélanges finis. Ce cadre nous permet de traiter le problème dans un contexte multivarié et fournit en outre plusieurs diagnostics utiles pour établir la priorité des cas qui doivent être examinés plus en profondeur par examen manuel. Le classement des unités par ordre de priorité est important si l'on veut accroître l'exactitude des données, tout en évitant de perdre du temps en faisant le suivi d'unités qui ne sont pas vraiment critiques.

Mots clés : Vérification; erreur aléatoire; erreur systématique; vérification sélective; classification fondée sur un modèle.

1. Introduction

Les éléments qui déterminent la qualité d'un processus de vérification et d'imputation (V et I) sont multiples et ont été décrits en détail dans la littérature (Granaquist 1995). Nous nous intéressons à une erreur non due à l'échantillonnage particulière qui a une forte incidence sur deux dimensions concurrentes importantes de la qualité, à savoir la rapidité de production et l'exactitude des données. En ce qui concerne l'exactitude, nous adoptons la définition proposée dans l'Encyclopedia of Statistical Sciences (1999) : [traduction] « L'exactitude s'entend de la concordance entre les statistiques et les caractéristiques cibles ». Un certain nombre de facteurs peuvent causer des inexactitudes tout au long du processus d'enquête statistique. L'inexactitude peut être réduite durant la phase de vérification et d'imputation, qu'on peut considérer comme un « outil d'amélioration de l'exactitude grâce auquel les données erronées ou très suspectes sont découvertes et, au besoin, corrigées (imputées) » (Federal Committee on Statistical Methodology

1990). Étant donné la complexité des phénomènes étudiés et l'existence de plusieurs types d'erreur non due à l'échantillonnage, la phase de vérification et d'imputation peut être très longue et complexe (Granaquist 1996). Dans la littérature spécialisée, une classification courante des erreurs repose sur la définition de deux catégories d'erreur, à savoir l'erreur systématique et l'erreur aléatoire. La première catégorie comprend les erreurs qui sont toutes de même signe et produisent un biais en statistique, tandis que la

seconde englobe les erreurs qui sont réparties aléatoirement autour de zéro et ont une incidence sur la variance des estimations (Encyclopedia of Statistical Sciences 1999). Comprendre la nature des erreurs aide non seulement à déterminer la source et à adopter la méthode convenant le mieux pour les corriger (Di Zio et Luzi 2002). Alors que l'approche de Fellegi-Holt (Fellegi et Holt 1976) est un modèle reconnu pour s'occuper des erreurs aléatoires, des solutions ponctuelles sont généralement adoptées pour traiter les erreurs systématiques (voir, par exemple, Euredit 2003, vol. 1, chapitre 5). Habituellement, les erreurs systématiques sont traitées avant les erreurs aléatoires, particulièrement si ces dernières le sont au moyen d'un logiciel automatisé, comme le Système généralisé de vérification et d'imputation (SGVI) (Kovar, Mac Millan et Whitridge 1988) et, plus récemment, De Waal (2003).

Dans la famille des erreurs systématiques, l'une dont l'incidence sur les estimations finales est importante et qui affecte fréquemment les données des enquêtes statistiques (par exemple, enquêtes auprès des entreprises) est l'erreur d'unité de mesure multipliée par une constante (par exemple, 100 ou 1 000). Cette erreur est due au fait que certains répondants choisissent incorrectement l'unité de mesure lors de la déclaration de la quantité de certains items du questionnaire.

À titre d'exemples d'enquêtes souffrant de ce genre d'erreur, nous avons choisi deux enquêtes réalisées par ISTAT, c'est-à-dire l'Enquête italienne sur le coût de la

Bibliographie

Abul-Elia, Abdel-Latif, A., Greenberg, B.G. et Horvitz, D.G. (1967). A multiproportions RR model. *Journal of the American Statistical Association*, 62, 990-1008.

Chaudhuri, A., et Mukerjee, R.M. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

Cochran, W.G. (1977). *Sampling Techniques*, 3^e ed. New York: John Wiley & Sons, Inc.

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W. et Ezziit, T.M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. Dans *Measurement Errors in Surveys* (Eds. P.P. Biemer et coll.). New York: John Wiley & Sons, Inc.

Droitcour, J.A., Larson, E.M. et Scheuren, F.J. (2001). The three card method: Estimating sensitive survey items with permanent anonymity of response. *Proceedings of the Social Statistics Section of the American Statistical Association*.

Greenberg, B.G., Abul-Elia, Abdel-Latif, A., Simmons, W.R. et Horvitz, D.G. (1969). The unrelated question RR model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

Hubbard, M.L., Casper, R.A. et Lessler, J.T. (1989). Respondent reactions to item count lists and randomized response. *Proceedings of the Survey Research Section of the American Statistical Association*, 544-548.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lessler, J.T., et O'Reilly J.M. (1997). Mode of interview and reporting sensitive issues: Design and implementation of audio computer-assisted self-interviewing. *NIDA Research Monograph*, 57, 104-124.

Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey: John Wiley & Sons, Inc.

Sakamoto, Y., Tsuchiya, T., Nakamura, T., Maeda, T. et Fouse, D.B. (2000). *A Study of the Japanese National Character: The Tenth Nationwide Survey* (1998). Tokyo: The Institute of Statistical Mathematics Research Report General Series 85.

Sämdal, C.-E., Swesson, B. et Weitzman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schuman, H., et Presser, S. (1981). *Questions & Answers in Attitude Surveys*. New York: Academic Press.

Takahasi, K., et Sakasegawa, H. (1977). A randomized response technique without making use of any randomizing device. *Annals of the Institute of Statistical Mathematics*, 29, 1-8.

U.S. General Accounting Office (1999). *Survey Methodology: An Innovative Technique for Estimating Sensitive Items*. Washington D.C.: General Accounting Office.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

Pour une question à laquelle les répondants évitent de donner une réponse honnête, la valeur réelle de $\pi = P(Y = 1)$ serait souvent faible. En outre, la méthode de questionnement utilisée est indirecte afin d'assurer la protection des renseignements personnels. Les répondants ont le sentiment que leur vie privée est protégée si un grand nombre d'items non clés sont inclus (Hubbard et coll., 1989). Les études en simulation montrent que, dans de telles situations, la méthode par croisement ou par double croisement est plus efficace que la méthode stratifiée classique. Les estimateurs de domaine obtenus par la méthode stratifiée classique ne convergent généralement pas vers l'estimateur $\hat{\pi}$ comme le montre l'équation (10). L'utilisation de l'estimateur $\hat{\pi}_{ps}$ poststratifié par la variable de domaine étudiée est essentielle pour assurer la convergence. Autrement, il faut diviser l'échantillon total en deux sous-groupes de façon telle que les distributions de leur variable de domaine concordent a priori. Par contre, les estimateurs de domaine obtenus par les méthodes par croisement et par double croisement convergent vers $\hat{\pi}$ tel que le montre l'équation (21). Cependant, cela ne signifie pas que la méthode par croisement donne automatiquement l'ajustement des deux sous-groupes de sorte que les distributions d'échantillon de la variable de domaine dans les deux sous-groupes concordent. Pour la méthode par croisement, la poststratification par les variables de domaine ou d'autres variables démographiques est acceptable, mais non indispensable.

Même si l'on utilise la méthode par double croisement, on observe parfois des estimations de domaine négatives. Il est possible d'éviter ces estimations négatives en permettant à une estimation négative q_c de Q_c dans (23) d'être nulle. Cependant, ce genre de correction produit un biais positif dans $p(Y = 1 | Z = z)$.

Les données de l'enquête sur le caractère national japonais, qui ont été utilisées pour les expériences en simulation, ne sont pas délicates et n'ont pas été obtenues par la technique du dénombrement d'items. Dans l'avenir, il faudrait évaluer les propriétés de la méthode proposée en l'appliquant à des données obtenues par cette technique.

Remerciements

L'auteur remercie deux examinateurs anonymes et le rédacteur adjoint de leurs commentaires constructifs au sujet d'une version antérieure du présent article.

Tableau 5
Moyenne et écart-type de e^2 et nombre d'estimations illogiques obtenues (la variable de domaine Z est le sexe)

Méthode stratifiée	Combinaison 1	Combinaison 2	Combinaison 3	Méthode par croisement			Méthode par double croisement			Réponse aléatoire		
	38	89	341	18	45	163	18	45	163	12	35	158
7 Original (7 %)												
Valeur e^2	(36)	(92)	(330)	(24)	(65)	(239)	(24)	(65)	(240)	(14)	(43)	(181)
inférieure	39	179	457	1	41	186	1	31	177	0	41	305
moyenne	0	0	0	0	0	0	0	0	0	0	0	0
supérieure	6	16	44	4	10	22	3	9	21	3	8	35
8 Poi (50 %)												
Valeur e^2	(6)	(17)	(43)	(5)	(12)	(31)	(4)	(12)	(31)	(3)	(7)	(34)
inférieure	0	0	0	0	0	0	0	0	0	0	0	0
moyenne	0	0	0	0	0	0	0	0	0	0	0	0
supérieure	0	0	0	0	0	0	0	0	0	0	0	0
2 Diligent (71 %)												
Valeur e^2	(4)	(11)	(32)	(3)	(8)	(23)	(2)	(3)	(23)	(2)	(5)	(23)
inférieure	0	0	0	0	0	0	0	0	0	0	0	0
moyenne	4	10	33	3	7	17	2	6	16	2	5	23
supérieure	0	0	0	0	0	0	0	0	0	0	0	0

Nota : La valeur e^2 est multipliée par 10³.

Tableau 6
Moyenne et écart-type de e^2 et nombre d'estimations illogiques obtenues (la variable de domaine Z est l'âge)

Méthode stratifiée	Combinaison 1	Combinaison 2	Combinaison 3	Méthode par croisement			Méthode par double croisement			Réponse aléatoire		
	375	859	3 410	93	175	536	70	153	526	158	476	2 181
7 Original (7 %)												
Valeur e^2	(226)	(507)	(2 108)	(82)	(195)	(733)	(75)	(202)	(745)	(101)	(294)	(1 348)
inférieure	609	799	926	8	138	273	8	93	246	284	720	945
moyenne	0	0	1	0	0	0	0	0	0	0	0	0
supérieure	60	152	446	32	80	89	13	45	72	25	74	335
8 Poi (50 %)												
Valeur e^2	(39)	(91)	(290)	(20)	(42)	(95)	(13)	(35)	(94)	(14)	(42)	(193)
inférieure	0	0	48	0	0	0	0	0	0	0	0	9
moyenne	0	0	41	0	0	0	0	0	0	0	0	9
supérieure	0	0	333	28	59	70	9	31	52	17	51	232
2 Diligent (71 %)												
Valeur e^2	(26)	(58)	(217)	(16)	(33)	(71)	(8)	(23)	(70)	(11)	(31)	(136)
inférieure	0	0	9	0	0	0	0	0	0	0	0	0
moyenne	0	0	9	0	0	0	0	0	0	0	0	0
supérieure	0	18	353	0	0	10	0	0	1	0	2	217

Nota : La valeur e^2 est multipliée par 10³.

4. Conclusion

Nos expériences en simulation ont produit les résultats suivants :

- La méthode par croisement ou la méthode par double croisement proposée dans le présent article devrait être utilisée pour estimer les paramètres de domaine lorsque les données sont obtenues par la technique du dénombrement d'items. Dans la première simulation, la variance des estimateurs par croisement était réduite à 39 % de la variance de

Même si l'on utilise la méthode par double croisement, l'erreur-type des estimateurs de domaine est beaucoup plus grande que celle produite par la technique de questionnement direct.

Nous utiliserons tous les échantillons comme suit :

Étape 1. Calculer $P(Y=1|Z=z)$ pour chaque z d'après toutes les données pour la taille $N=1\,339$.

Étape 2. Diviser l'échantillon total ($N=1\,339$) aléatoirement en un groupe A et un groupe B de taille n^A et n^B où $N=n^A+n^B$.

Étape 3. Compter le nombre C^A d'items non clés sélectionnés pour chaque répondant du groupe A et compter le nombre C^B d'items sélectionnés, y compte l'item clé et les items non clés, dans le groupe B.

Étape 4. Estimer $p(Y=1|Z=z)$ par la méthode stratifiée, la méthode par croisement et la méthode par double croisement, respectivement.

Étape 5. Calculer la distance du chi-carré e^2 entre $P(Y=1|Z=z)$ et $p(Y=1|Z=z)$ pour chaque méthode.

$$e^2 = \sum^z \frac{p(Y=1|Z=z) - P(Y=1|Z=z)}{p(Y=1|Z=z)^2}$$

Étape 6. Répéter la procédure susmentionnée de l'étape 2 à l'étape 5 pour 1 000 itérations. Calculer la moyenne et l'écart-type de e^2 pour chaque méthode.

En outre, à titre de référence, nous avons simulé la méthode stratifiée sous réponse aléatoire par la procédure suivante :

Étape 1. Soit p une proportion telle que décrite plus loin. Diviser l'échantillon total ($N=1\,339$) aléatoirement en deux groupes. Le groupe A est composé de Np répondants et le groupe B, de $N(1-p)$ répondants.

Étape 2. Soit n^A le nombre de répondants qui ont choisi l'item clé et $Z=z$ dans le groupe A. Soit n^B le nombre de répondants qui n'ont pas choisi l'item clé et $Z=z$ dans le groupe B. Soit n^z le nombre de répondants avec $Z=z$. Calculer

$$p(Y=1|Z=z) = \frac{1}{n^z} \left(\frac{1}{1.339} \left(p - 1 + \frac{n^A}{n^z} + \frac{n^B}{n^z} \right) \right)$$

Étape 3. Calculer e^2 en se servant de la même équation que celle utilisée pour la technique du dénombrement d'items.

3.3.2 Résultats des simulations

Nous avons utilisé trois valeurs de p , à savoir $p=0,2$, $p=0,3$ et $p=0,4$.

Étape 4. Répéter la procédure susmentionnée de l'étape 1 à l'étape 3 pour 1 000 itérations. Calculer la moyenne et l'écart-type de e^2 pour chaque méthode.

Les tableaux 5 et 6 donnent la moyenne et l'écart-type pour 1 000 e^2 pour la variable de domaine Z de sexe et d'âge, respectivement. Les estimateurs de domaine sont d'autant plus précis que la moyenne de la « valeur e^2 » est faible. Pour certaines répétitions, nous avons obtenu des estimations illogiques de $p(Y=1|Z=z)$, qui s'écartent de l'intervalle [0, 1]. Les colonnes des tableaux intitulées « Inférieure » indiquent le nombre de répétitions pour lesquelles au moins une des estimations de $p(Y=1|Z=z)$ était inférieure à 0 et les colonnes intitulées « Supérieure » indiquent le nombre d'estimations qui étaient supérieures à 1. Idéalement, les chiffres figurant dans les colonnes « p illogique » devraient être nuls.

Pour toute combinaison de l'item clé, des items non clés et de la variable de domaine Z , les moyennes de e^2 calculées pour la méthode par double croisement sont les plus faibles et celles pour la méthode par croisement viennent au deuxième rang, l'écart étant très faible. Quand la proportion π de l'item clé est faible (« 7 Original »), que le nombre d'items non clés est grand (combinaison 3) et que le nombre d'options de la variable de domaine Z est grand (âge), la précision de la méthode stratifiée diminue considérablement comparativement aux autres combinaisons.

En outre, quand la proportion π de l'item clé est faible, l'utilisation de la méthode stratifiée produit souvent des estimations négatives. Par exemple, si l'on combine « 7 Original », la combinaison 3 et l'âge, la fréquence observée des estimations négatives est de 926 pour 1 000 itérations. Si l'on utilise la méthode par double croisement, les estimations négatives sont moins fréquentes. Pour la méthode de la réponse aléatoire, si le nombre d'options de la variable de domaine Z est faible (sexe), les estimations semblent avoir la même précision que celles obtenues par les méthodes par croisement et par double croisement. Cependant, la moyenne de e^2 est un peu plus grande que celle observée pour la méthode par croisement quand le nombre d'options pour la variable de domaine Z est grand (âge). La méthode de la réponse aléatoire, pour laquelle seule la méthode stratifiée est disponible, produit aussi des estimations négatives, surtout quand π est faible (« 7 Original »).

dénombrement d'items est près de 7 à 26 fois plus élevée que pour la méthode de questionnement direct. Cependant, la réduction de la variance due à l'utilisation de la méthode par double croisement au lieu de la méthode stratifiée varie de $Def_{w,s} = 0,39$ (hommes) à 0,55 (femmes). Donc, l'erreur-type de la méthode par double croisement correspond à environ 62 % de celle de la méthode stratifiée, pour la valeur minimale et à 74 %, pour la valeur maximale.

3.3 Méthode stratifiée contre méthode par croisement

3.3.1 Méthodes de simulation

À la section précédente, nous avons montré que la précision des méthodes par croisement et par double croisement semble être plus grande que celle de la méthode stratifiée. Nous allons maintenant vérifier la précision de ces méthodes pour d'autres combinaisons de l'item clé, de la combinaison d'items non clés et de la variable de domaine Z grâce à des expériences en simulation.

Afin d'évaluer le degré de variance ou l'erreur-type des estimateurs, considérons le critère suivant, qui est analogue à l'effet du plan (Kish 1965),

$$Def_{M_1, M_2} = \frac{SE_2^2}{SE_1^2},$$

où M_1 et M_2 indiquent l'une des quatre méthodes D , S , C et W . Nous ne présentons pas les résultats détaillés, mais brièvement, $Def_{C,D}$ varie de 50 (quand $f = 0,1$) à 700 (quand $f = 0,9$). Autrement dit, même si nous utilisons la méthode par croisement, l'erreur-type de la technique du

laquelle les erreurs-types ne convergent pas vers zéro même si $f = 1$. Comme nous l'avons mentionné plus haut, cette situation est due à la randomisation introduite à la phase de subdivision de l'échantillon. Dans le cas de la méthode stratifiée, les erreurs-types sont manifestement plus grandes que pour les deux méthodes par croisement. Les courbes produites pour les méthodes par croisement et par double croisement se superposent pour ainsi dire et ne présentent aucune différence marquée.

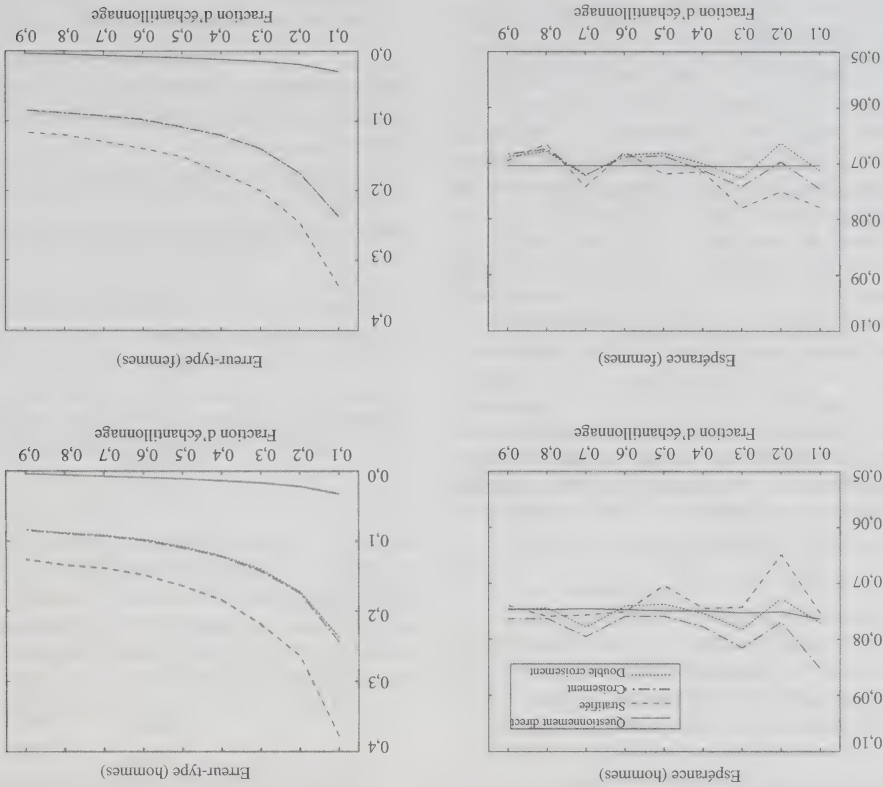


Figure 1. Espérance et erreur-type approximative des estimateurs.

Nous utilisons trois combinaisons d'items non clés, telles qu'énumérées au tableau 3. La combinaison 1 comprend deux items pour lesquels la proportion est faible, tandis que la combinaison 2 comprend deux items dont la proportion est élevée. La combinaison 3 est le cas où le nombre d'items non clés est maximal.

Tableau 3

Trois combinaisons d'items non clés	
Items non clés	
Combinaison 1 (G = 2):	9 Joyeux (8%)
	3 Libre (13%)
Combinaison 2 (G = 2):	5 Persistant (51%)
	6 Gentil (42%)
Combinaison 3 (G = 2):	Neuf items autres que l'item clé

Nous utilisons soit le sexe, soit l'âge comme variable de domaine Z. Le sexe est masculin ou féminin, et les groupes d'âges sont « 20 à 29 ans », « 30 à 39 ans », « 40 à 49 ans », « 50 à 59 ans », « 60 à 69 ans » et « 70 ans et plus ».

3.2 Questionnement direct contre technique du dénombrement d'items

3.2.1 Méthodes de simulation

Premièrement, nous comparons les erreurs-types des techniques de questionnement direct et de dénombrement d'items. Dans cette expérience, nous avons testé la combinaison de « 7 Original » (item clé), la combinaison 3 (items non clés) et le sexe (variable de domaine). Le tableau de contingence fondé sur l'échantillon complet de $N = 1\,339$ figure au tableau 4.

Tableau 4

Tableau de contingence entre « 7 Original » et le sexe

7 Original	
Y = 1	Y = 0
Hommes 46 (7,5) 569 (92,5)	615 (100,0) 724 (100,0)
Femmes 51 (7,0) 673 (93,0)	1 339 (100,0) 1 242 (92,8)
Total 97 (7,2)	1 242 (92,8)

La simulation a été réalisée selon la procédure suivante :

Étape 1. Poser que l'échantillon total de $N = 1\,339$ est une population.

Étape 2. Tirer un sous-échantillon S de taille N_f , où f est la fraction d'échantillonnage sous échantillonnage aléatoire simple sans remise.

Étape 3. À titre de résultat simulé de la méthode de questionnement direct, calculer directement les proportions $p(Y = 1|Z = \text{hommes})$ et $p(Y = 1|Z = \text{femmes})$.

Étape 4. Diviser le sous-échantillon S en deux groupes S^A et S^B de tailles n^A et n^B qui ne sont pas nécessairement égales. Compter le nombre C^A d'éléments non clés sélectionnés pour chaque répondant dans S^A . En outre, compter le nombre C^B d'éléments sélectionnés, y compris l'élément clé et les éléments non clés, dans S^B .

Étape 5. À titre de résultat simulé de la technique de dénombrement d'items, calculer $p(Y = 1|Z = \text{hommes})$ et $p(Y = 1|Z = \text{femmes})$ par les trois méthodes, à savoir la méthode stratifiée, la méthode par croisement et la méthode par double croisement. Dans la méthode par double croisement, poser que $w^A = n^A/(n^A + n^B)$ et $w^B = n^B/(n^A + n^B)$.

Étape 6. Poser que $f = 0,1$ à l'étape 2 et exécuter les étapes 2 à 5 pour 2 000 itérations. Calculer les moyennes $E_{D_p}, E_{S_p}, E_{C_p}$ et les écarts-types $SE_{D_p}, SE_{S_p}, SE_{C_p}$ pour chaque méthode d'estimation afin d'obtenir une approximation des espérances et des erreurs-types des estimateurs, où les indices D_p, S_p, C_p et W indiquent la méthode de questionnement direct, la méthode stratifiée, la méthode par croisement et la méthode par double croisement, respectivement. De la même façon, poser que $f = 0,2$ et exécuter les étapes 2 à 5 pour 2 000 itérations, et ainsi de suite jusqu'à $f = 0,9$ inclusivement.

3.2 Résultats des simulations

La figure 1 donne les espérances et les erreurs-types approuvées des estimateurs. Les axes horizontaux donnent la fraction d'échantillonnage f . Dans les deux cas, c'est-à-dire les hommes et les femmes, l'espérance approximative de E_{D_p} est stable pour les diverses valeurs de f , tandis que les espérances de E_{S_p}, E_{C_p} et E_W pour la technique du dénombrement d'items fluctuent irrégulièrement. Ces fluctuations sont dues au fait que la méthode du dénombrement d'items comprend deux randomisations, à la phase d'échantillonnage et à la phase de la subdivision de l'échantillon, tandis que le scénario de questionnement direct ne comprend qu'une seule randomisation à la phase d'échantillonnage. Même si $f = 1$, l'estimateur sous la technique du dénombrement direct présente une certaine variance due à la randomisation à la phase de la subdivision de l'échantillon. Comme l'étendue des fluctuations est négligeable comparativement à la grandeur des erreurs-types, illustrées plus bas, nous concluons que le nombre de répétitions était suffisant.

Les erreurs-types, SE_{D_p} , pour la méthode de questionnement direct sont nettement plus faibles que celles observées pour la technique du dénombrement d'items, pour

$$\sum_{Z=1}^Z P(Y=1, Z=z) = \sum_{G=1}^G \sum_{C=1}^C \sum_{A=1}^A \left(\frac{n_A^D}{n_A^P} - \frac{n_B^D}{n_B^P} \right) \left(\frac{n_A^D}{n_A^P} - \frac{n_B^D}{n_B^P} \right) \left(1 - \sum_{G=1}^G \frac{n_A^D}{n_A^P} \right) - \left(1 - \sum_{G=1}^G \frac{n_B^D}{n_B^P} \right)$$

$$= \sum_{G=1}^G \sum_{C=1}^C \frac{n_B^D}{n_B^P} - \sum_{G=1}^G \sum_{C=1}^C \frac{n_A^D}{n_A^P} = \pi. \tag{21}$$

Naturellement, si nous connaissons les proportions de domaine $P(Z=z)=m_z$, nous pouvons les utiliser pour obtenir un estimateur poststratifié $p(C^A=d)$ dans \mathcal{Q}^{-1} de (14),

$$p(C^A=d) = \sum_{z=1}^z \frac{m_z^2}{m_z^2 n_{dz}^2}.$$

Dans ce cas, $\sum_z p(Y=1, Z=z)$ coïncide avec l'estimateur poststratifié π_{ps} .

Un inconvénient de la méthode par croisement est que la variance de $p(Y=1|Z=z)$ est presque impossible à estimer algébriquement. Par conséquent, il faut utiliser une méthode par rééchantillonnage, telle que le jackknife ou le bootstrap. De plus, puisqu'il est impossible de déterminer laquelle, de la méthode stratifiée et de la méthode par croisement, est la plus efficace, nous décrivons plus loin la réalisation d'une étude en simulation.

2.3 Méthode par double croisement

Avant de passer à l'étude en simulation, nous proposons une version modifiée de la méthode par croisement. Dans l'équation (19) de la méthode par croisement, nous utilisons $p(Z=z|C^B=c)$. De la même façon, l'utilisation de $p(Z=z|C^A=c)$ donne

$$p(Y=1, Z=z) = \sum_{G=1}^G p(Z=z|C^A=c) p(C^A=c, Y=1)$$

$$= \sum_{G=1}^G p(Z=z|C^A=c) \tilde{Q}_c. \tag{22}$$

Par conséquent, nous obtenons une méthode par double croisement en combinant (14) et (22) comme suit :

$$p(Y=1, Z=z) = \sum_{G=1}^G \left\{ w_A^P(Z=z|C^A=c) + w_B^P(Z=z|C^B=c+1) \right\} \tilde{Q}_c, \tag{23}$$

où w_A^P et w_B^P sont les poids non négatifs de chaque sous-groupe, dont la somme est égale à 1.

L'équation qui suit est également vraie pour la méthode par double croisement de tout poids w_A^P et w_B^P , à moins que $n_B^A=0$ ou $n_B^B=0$ pour certaines valeurs de c .

3. Expériences numériques

3.1 Ensemble de données

Afin de comparer la précision des estimateurs, nous avons réalisé des expériences en simulation en nous servant de données tirées de l'enquête sur le caractère national japonais (Sakamoto, Tsuchiya, Nakamura, Maeda et Fouse, 2000). Bien que les répondants aient été sélectionnés par échantillonnage stratifié à deux degrés parmi la population du Japon de 20 ans et plus, nous n'avons pas tenu compte du plan d'échantillonnage, parce que nous avons traité l'échantillon recueilli de $N=1\,339$ comme étant la population « réelle » dans cette expérience. Le tableau 2 donne les résultats pour une question au sujet des attributs significatifs du caractère japonais. Lors d'une interview sur place, on a demandé aux répondants de choisir, parmi une liste de dix adjectifs, tous ceux qui, selon eux, décrivaient le caractère japonais.

Tableau 2

Attributs significatifs du caractère japonais

N = 1 339	
(Montrer la carte) Selon vous, lesquels, parmi les adjectifs suivants (décrire le caractère du peuple japonais? Choisissez autant d'adjectifs que vous souhaitez.	
1 Rationnel	18 %
2 Diligent	71 %
3 Libré	13 %
4 Ouvert, franc	14 %
5 Persistant	51 %
6 Gentil	18 %
7 Original	71 %
8 Poli	8 %
9 Joyeux	14 %
10 Idéaliste	51 %
42 %	
7 %	
50 %	
8 %	
23 %	

La forme de cette question diffère de celle de la technique du dénombrement d'items, qui consiste à demander aux répondants d'« indiquer le nombre d'adjectifs ». Dans l'enquête décrite ici, on demande aux répondants d'« encercler autant d'adjectifs qu'ils jugent appropriés ». En outre, les dix items ne sont pas de nature très délicate, si bien que les répondants ne devaient pas hésiter durant la sélection. Cependant, puisque nous obtenons les tableaux de contingence réels entre chacun des dix éléments et une autre variable Z, nous pouvons évaluer les propriétés des estimateurs grâce à une pseudo procédure de dénombrement d'items.

Nous avons choisi chacun des trois items suivants comme item clé Y , où $Y=1$ signifie que l'item a été sélectionné.

- 7 Original (π la plus faible parmi les dix items)
- 8 Poli (π exactement égale à 50 %)
- 2 Diligent (π la plus grande parmi les dix items)

nous commençons par estimer la proportion conjointe $P(Y = 1, Z = z)$ afin d'utiliser l'échantillon complet, puis nous obtenons la proportion conditionnelle par

$$P(Y = 1 | Z = z) = \frac{P(Y = 1, Z = z)}{P(Z = z)} \quad \text{ou} \quad P(Y = 1 | Z = z) = \frac{P(Y = 1, Z = z)}{P(Z = z)}$$

Nous utilisons la dénomination « méthode par croisement », parce que cette méthode s'appuie sur des totalisations croisées $P(Z = z | C^B = c)$, comme le montre (19).

Dans la méthode par croisement, nous supposons que les équations qui suivent tiennent pour chaque valeur de c .

Hypothèse 2.

$$(11) \quad P(C^B = c + 1, Y = 1) = P(C^A = c, Y = 1),$$

$$(12) \quad P(C^B = 0, Y = 1) = P(C^A = -1, Y = 1) = 0,$$

$$(13) \quad P(C^B = c, Y = 0) = P(C^A = c, Y = 0).$$

Ces hypothèses impliquent aussi que la différence entre les distributions de C^A et C^B dépend uniquement de Y .

En nous fondant sur ces hypothèses, nous obtenons le résultat suivant.

Méthode par croisement.

$$(14) \quad P(Y = 1, Z = z) = z \sum_{G=1}^c P(Z = z | C^B = c) Q^{c-1},$$

où

$$Q^c = \sum_{c=1}^z \{P(C^A = d) - P(C^B = d)\}.$$

De plus, on suppose que $P(Z = z | C^B = c, Y = 1) = P(Z = z | C^B = c)$ pour tout $c > 0$. Cette hypothèse serait valide jusqu'à un certain point, quand les items clés et non clés décrivent tous les deux le même type de comportement stigmatisant.

Calculs.

D'après les hypothèses, nous avons

$$P(C^B = c) = P(C^B = c, Y = 1) + P(C^B = c, Y = 0)$$

$$(15) \quad = P(C^A = c - 1, Y = 1) + P(C^A = c, Y = 0).$$

L'équation qui suit est vérifiée pour toute valeur de c .

$$(16) \quad P(C^A = c, Y = 0) = P(C^A = c) - P(C^A = c, Y = 1).$$

Donc, la substitution de (16) dans (15) donne

$$P(C^B = c) = P(C^A = c - 1, Y = 1)$$

$$(17) \quad + \{P(C^A = c) - P(C^A = c, Y = 1)\}.$$

Par sommation de (17) sur c jusqu'à une certaine valeur g ,

$$\sum_{g=0}^c P(C^B = c) = \sum_{g=0}^c P(C^A = c - 1, Y = 1)$$

$$+ \sum_{g=0}^c \{P(C^A = c) - P(C^A = c, Y = 1)\}$$

$$= \sum_{g=0}^c P(C^A = c) - P(C^A = g, Y = 1).$$

Par transposition des termes, nous définissons \tilde{Q}^c .

$$\tilde{Q}^c = \sum_{c=0}^p \{P(C^A = d) - P(C^B = d)\}$$

$$= P(C^A = c, Y = 1)$$

$$= P(C^B = c + 1, Y = 1).$$

(18)

Ici, la proportion conjointe $P(Y = 1, Z = z)$ se décompose comme suit

$$P(Y = 1, Z = z) = \sum_{G=1}^c P(Z = z | C^B = c) P(C^B = c, Y = 1). \quad (19)$$

La substitution de l'équation (18) et de l'hypothèse (12)

dans l'équation (19) donne la méthode par croisement. L'estimateur conjoint $P(Y = 1, Z = z)$ s'obtient par remplacement de chaque terme de (14) par son estimateur.

Si l'échantillon est autopondéré, l'estimateur s'écrit

$$(20) \quad P(Y = 1, Z = z) = \sum_{G=1}^c \sum_{c=1}^p \frac{n_B^{n_{cz}}}{n_B^{n_{cz}}} \left(\frac{n_A^{n_A}}{n_B^{n_A}} - \frac{n_B^{n_B}}{n_B^{n_B}} \right),$$

où

$$\sum_{c=1}^z n_A^{n_{cz}} = \sum_{c=1}^z n_A^{n_{cz}} \quad \text{et} \quad \sum_{c=1}^z n_B^{n_{cz}} = \sum_{c=1}^z n_B^{n_{cz}}.$$

Nous obtenons l'estimateur conditionnel $P(Y = 1 | Z = z)$ en divisant $P(Y = 1 | Z = z)$ par les proportions de domaine $P(Z = z)$ ou leur estimateur $P(Z = z)$.

Comme nous l'avons mentionné plus haut, la caractéristique principale de la méthode par croisement est qu'on commence par estimer $P(Y = 1, Z = z)$ pour l'échantillon complet. Par conséquent, la variance de

$P(Y = 1 | Z = z)$ devrait être plus faible dans le cas de la

méthode par croisement que dans celui de la méthode stratifiée. En outre, la méthode par croisement produit

rarement des valeurs négatives, tandis que la méthode stratifiée en produit fréquemment. De surcroît, l'estimateur

marginal $P(Y = 1)$ obtenu par sommation de (20) est égal à l'estimateur (3), à moins que $n_B^c = 0$ pour certaines valeurs de c :

2. Estimateurs de domaine pour la technique du dénombrement d'items

2.1 Méthode stratifiée

Ici, nous reformulons la méthode stratifiée. Supposons que les équations qui suivent soient vérifiées pour chaque valeur de c et de z .

Hypothèse 1.

$$P(C_B^z = c | Z = z) = P(C_A^z = c, Y = 0 | Z = z) + P(C_A^z = c - 1, Y = 1 | Z = z) + P(C_A^z = G + 1, Y = 0 | Z = z) = 0.$$

Ces hypothèses sous-entendent que la différence entre les distributions de C_A^z et C_B^z dépend uniquement de Y . Les effets de question, dont les effets d'ordre et les effets de contexte (Schuman et Presser 1981), ne sont pas pris en considération.

En nous fondant sur ces hypothèses, nous obtenons le résultat suivant.

Méthode stratifiée.

$$P(Y = 1 | Z = z) = \sum_{G+1}^G P(C_B^z = c | Z = z) - \sum_{G}^0 P(C_A^z = c | Z = z)$$

$$(5) \quad = \underline{C}_B^z - \underline{C}_A^z, \\ (6)$$

où \underline{C}_A^z et \underline{C}_B^z sont les moyennes de domaine de C_A^z et C_B^z .

Calculs.

$$\begin{aligned} \sum_{G+1}^G P(C_B^z = c | Z = z) &= \sum_{G+1}^G P(C_A^z = c, Y = 0 | Z = z) + \sum_{G+1}^G P(C_A^z = c - 1, Y = 1 | Z = z) \\ &= \sum_{G}^0 P(C_A^z = c, Y = 0 | Z = z) + \sum_{G}^0 (c + 1) P(C_A^z = c, Y = 1 | Z = z) \\ &= \sum_{G}^0 c \{ P(C_A^z = c, Y = 0 | Z = z) + P(C_A^z = c, Y = 1 | Z = z) \} \\ &+ \sum_{G}^0 P(C_A^z = c, Y = 1 | Z = z) \\ &= \sum_{G}^0 c P(C_A^z = c | Z = z) + P(Y = 1 | Z = z). \end{aligned}$$

Le transfert du premier terme dans le premier membre de l'équation donne la méthode stratifiée (5). L'estimateur $P(Y = 1 | Z = z)$ s'obtient en remplaçant les moyennes de domaine \underline{C}_A^z et \underline{C}_B^z par leurs estimateurs, \hat{C}_A^z et \hat{C}_B^z .

Quand les probabilités de sélection sont égales pour toutes les unités de l'échantillon, l'estimateur $P(Y = 1 | Z = z)$ s'écrit

$$(7) \quad P(Y = 1 | Z = z) = \sum_{G+1}^G c \frac{n_B^z}{n_B} - \sum_{G}^0 c \frac{n_A^z}{n_A},$$

où n_A^z, n_B^z, n_A^{cz} et n_B^{cz} sont définis à la section 1.1. Les équations (2) et (3) pour l'ensemble de la population sont des cas particuliers de (7) et (8).

L'un des avantages de la méthode stratifiée est que l'estimateur de la variance de $P(Y = 1 | Z = z)$ s'obtient facilement par

$$(9) \quad \text{Var}(P(Y = 1 | Z = z)) = \text{Var}(\hat{C}_B^z) + \text{Var}(\hat{C}_A^z).$$

Par ailleurs, comme nous l'avons souligné à la section précédente, la réduction de la taille d'échantillon dans chaque strate augmente les variances estimées dans (9). De surcroît, l'estimateur marginal $P(Y = 1)$ obtenu en utilisant (8) ne correspond pas à celui obtenu directement au moyen de (3), à moins que $n_A^z = n_B^z$ pour tout z . Autrement dit, si $P(Z = z)$ n'est pas connue, son estimateur est donné par

$$\text{et} \quad \sum_{G+1}^G P(Y = 1 | Z = z) P(Z = z) = (n_A^z + n_B^z) / (n_A^z + n_B^z)$$

$$= \sum_{G+1}^G \frac{n_A^z + n_B^z}{n_A^z + n_B^z} \left\{ \sum_{G+1}^G c \frac{n_B^z}{n_B} - \sum_{G}^0 c \frac{n_A^z}{n_A} \right\} \\ \neq \sum_{G+1}^G c \frac{n_B^z}{n_B} - \sum_{G}^0 c \frac{n_A^z}{n_A} = \hat{\pi}.$$

Si l'on connaît la proportion de domaine $P(Z = z) = m_z$, l'estimateur marginal correspond à l'estimateur post-stratifié (4).

$$\sum_{G+1}^G P(Y = 1 | Z = z) P(Z = z) = \sum_{G+1}^G m_z \left\{ \sum_{G+1}^G c \frac{n_B^z}{n_B} - \sum_{G}^0 c \frac{n_A^z}{n_A} \right\} = \hat{\pi}_{\text{ps}}.$$

Ces résultats indiquent que nous devrions utiliser un estimateur poststratifié $\hat{\pi}_{\text{ps}}$ avec les estimateurs de domaine si nous utilisons la méthode stratifiée.

2.2 Méthode par croisement

Dans la méthode stratifiée, nous subdivisons un échantillon de l'ensemble de la population en strates afin de procéder à l'estimation directe de $P(Y = 1 | Z = z)$, ce qui cause une réduction de la taille d'échantillon. Par conséquent, dans la méthode par croisement proposée ici,

où n_A^c est le nombre de répondants pour chaque $C^A = c$ et $Z = z$,

$$n_A^c = \sum_{G=0}^z n_A^{cz}, n_A^c = \sum_{z=0}^Z n_A^{cz}, n_A^c = \sum_{z=0}^Z n_A^{cz}$$

et n_B^{cz}, n_B^z, n_B^c sont définis de façon analogue.

L'un des avantages de la technique du dénombrement d'items est qu'elle ne nécessite aucun des mécanismes de randomisation utilisés dans la technique de la réponse aléatoire. Ce n'est pas le répondant, mais un chercheur, qui choisit le questionnaire auquel il faut répondre. Donc, la technique du dénombrement d'items est facile à appliquer au moyen d'une enquête avec questionnaire à remplir soi-même ou d'une enquête téléphonique. Une comparaison plus approfondie de la technique de la réponse aléatoire et de la technique du dénombrement d'items figure dans Hubbard, Casper et Lessler (1989).

Le questionnaire A est introduit pour obtenir la distribution du nombre d'items non clés. Autrement dit, les personnes qui répondent au questionnaire A ne répondent pas à la question délicate. Par conséquent, il est possible d'accroître la précision de l'estimateur en utilisant la version à liste double de la technique du dénombrement d'items (Droitcour et coll., 1991), en vertu de laquelle il y a échange de rôle entre les deux sous-groupes. Cependant, ici, nous limitons notre argument à une version à liste unique, parce que l'extension des estimateurs à la version à liste double est simple.

1.2 But du présent article

Jusqu'à présent, nous sommes concentrés sur le paramètre $\pi = P(X=1)$ de l'ensemble d'une population. Cependant, il est souvent nécessaire d'obtenir des estimateurs pour des sous-populations ou domaines (Särndal, Swenson et Wretman 1992, page 5), c'est-à-dire d'estimer une proportion conditionnelle $P(X=1|Z=z)$ ou une proportion conjointe $P(X=1, Z=z)$, où la population est subdivisée en plusieurs domaines par la valeur Z. Ici, nous donnons à la variable Z le nom de variable de domaine. Les variables de domaine souvent utilisées sont les caractéristiques démographiques, comme le sexe ou l'âge. Par exemple, certains organismes gouvernementaux voudraient connaître, pour chaque groupe d'âge, la proportion de personnes qui consomment une drogue illicite particulière. Même si, dans l'équation (4), l'estimateur poststratifié $\hat{\pi}_{ps}$ utilise la variable de domaine Z, le but est d'estimer $P(X=1)$ pour l'ensemble de la population. Notre objectif, dans le présent article, est d'obtenir des estimations distinctes de $P(X=1|Z=z)$ pour chaque domaine. Voici une méthode simple d'estimation :

1. Poststratifier l'échantillon en strates ou domaines en se basant sur la valeur Z.

2. Dans chaque strate ou domaine, déterminer séparément $p(X=1|Z=z)$ en se servant de (1) ou (2), où $p(\cdot)$ est une estimation échantillonnale de $P(\cdot)$.
3. Au besoin, estimer $p(X=1, Z=z)$ en multipliant une proportion de domaine connue, $P(Z=z)$, ou une proportion de domaine estimée, $p(Z=z)$.

Dans tout l'article, nous appelons la méthode susmentionnée méthode stratifiée, parce que les estimations sont obtenues séparément pour chaque strate ou domaine. Bien que Rao (2003) donne à cette méthode le nom d'estimation directe, nous avons choisi de ne pas utiliser le terme « directe » afin d'éviter toute confusion avec l'expression « technique de questionnement direct ».

L'un des avantages de la méthode stratifiée est qu'elle s'applique à toute technique de questionnement indirect, y compris celles de la réponse aléatoire et du dénombrement d'items. Aux États-Unis, le General Accounting Office (1999) a adopté la méthode stratifiée pour produire des estimations de domaine par la technique des trois cartes. Cependant, l'un des inconvénients sérieux de la méthode stratifiée est qu'elle produit souvent des estimations illogiques, surtout des estimations négatives, dans le cas de la réponse aléatoire et du dénombrement d'items, comme nous l'expliquons plus loin. Ce problème tient principalement au fait que la réduction de la taille d'échantillon dans chaque strate accroît l'erreur-type des estimateurs (Lessler et O'Reilly 1997). Ainsi, Droitcour et coll. (1991, page 206) ont calculé des estimations distinctes pour les trois strates de risque et obtenu des estimations négatives du taux de prévalence de la consommation de drogue.

Dans le cas de la technique de la réponse aléatoire, les possibilités de produire des estimations de domaine autre-ment que par la méthode stratifiée sont peu nombreuses, parce que l'information sur le type de questionnaire choisi par les répondants n'est pas disponible. Par contre, dans le cas du dénombrement d'items, on sait à quel questionnaire a répondu chaque personne. Par conséquent, la précision des estimateurs de domaine devrait, en principe, augmenter si l'on utilise des données auxiliaires, plus précisément des tableaux de contingence entre Z et C^A ou C^B .

Dans le présent article, nous proposons pour la technique du dénombrement d'items, de nouveaux estimateurs de domaine que nous appelons méthode par croisement et méthode par double croisement, respectivement. En outre, nous montrons que les nouveaux estimateurs sont plus efficaces que la méthode stratifiée classique en simulant la technique du dénombrement d'items au moyen de données tirées de l'enquête sur le caractère national japonais visant à déterminer les attributs significatifs du caractère japonais.

chercheurs ne peuvent déterminer quels items sont vrais pour le répondant. Ainsi, un répondant pourrait indiquer que quatre items du questionnaire B sont vrais, mais nous ne pouvons être certains qu'il consomme de la drogue. Par conséquent, on devrait s'attendre à ce qu'un plus grand nombre de répondants consommant une drogue illícite donnent une réponse honnête dans une telle situation que si on leur posait une question directe.

3. Diviser l'échantillon en deux sous-groupes, A et B , de taille n_A et n_B au hasard, de sorte que chaque questionnaire soit attribué à un sous-groupe particulier.

Tableau 1

Exemples de listes d'items	
Questionnaire A	Questionnaire B
Combien parmi les items suivants sont vrais pour vous?	Combien parmi les items suivants sont vrais pour vous?
- posséder une bicyclette	- posséder une bicyclette
- avoir voyagé à l'étranger	- avoir voyagé à l'étranger
- avoir appelé une ambulance	- avoir appelé une ambulance
- posséder une villa d'été	- posséder une villa d'été
	illícite particulière
	- avoir consommé une drogue
	- posséder une villa d'été

« Appartenez-vous à $U^{(T)}$? ». Soit p ($\neq 0,5$) la probabilité prédéterminée. Chaque répondant choisit le questionnaire A ou B avec la probabilité p ou $1 - p$ respectivement, mais personne d'autre que le répondant ne sait quel questionnaire il a sélectionné.

3. Supposons que X est une variable indicateur dont la valeur est 1 si la réponse est « oui » ou 0 si la réponse est « non ». L'estimateur de π est donné par

$$\hat{\pi} = \frac{2p - 1}{p - 1 + X}, \tag{1}$$

où \bar{x} est une moyenne d'échantillon de X .

Puisque les chercheurs n'ont aucune information quant au type de questionnaire choisi par chaque répondant, un plus grand nombre de répondants devrait, en principe, répondre honnêtement que si on leur posait directement les questions.

La technique du dénombrement d'items, qui est le sujet du présent article, n'est pas aussi répandue malgré sa simplicité. Cette technique s'avère, elle aussi, efficace lorsqu'on doit poser des questions délicates, parce qu'on demande aux répondants non pas de répondre directement à ces questions, mais plutôt de déclarer le nombre d'items qui, parmi une liste, sont vrais dans leur cas. Les procédures de la technique du dénombrement d'items sont les suivantes :

1. Préparer l'item clé T , qui est le point de concentration de l'étude, et G autres items non clés E_1, \dots, E_G . Par exemple, T pourrait être « consommer une certaine drogue illícite », comme mentionné plus haut, et E_g , une description non délicate quelconque, telle que « posséder une bicyclette ».
2. Préparer deux types de questionnaires, A et B . Dans le questionnaire A , on demande aux répondants d'indiquer le nombre C_A d'items qui sont vrais en ce questionnaire B , on demande aux répondants d'indiquer le nombre C_B d'items qui sont vrais en ce questionnaire B et les quatre autres items ne sont pas des items clés. Sauf si une réponse au questionnaire B est $C_B = 0$ ou $C_B = 5$, les

Le tableau 1 donne des exemples de listes d'items. Notre but est d'estimer la proportion de personnes qui utilisent une drogue illícite particulière. L'item clé est « avoir consommé une drogue illícite particulière » dans le questionnaire B et les quatre autres items ne sont pas des items clés. Sauf si une réponse au questionnaire B est $C_B = 0$ ou $C_B = 5$, les

4. L'estimateur de π est donné par
- $$\hat{\pi} = \hat{C}_B - \hat{C}_A, \tag{2}$$
- où \hat{C}_A et \hat{C}_B sont les moyennes estimées de C_A et C_B , respectivement. La justification de (2) est donnée à la section 2.1. Si la probabilité de sélection est la même pour toutes les unités de l'échantillon, $\hat{\pi}$ peut s'écrire
- $$\hat{\pi} = \sum_{G+1}^{\infty} c \frac{n_B}{n_A} - \sum_G^{\infty} c \frac{n_A}{n_A}, \tag{3}$$

2. Préparer deux types de questionnaires, A et B . Dans le questionnaire A , on demande aux répondants d'indiquer le nombre C_A d'items qui sont vrais en ce questionnaire B , on demande aux répondants d'indiquer le nombre C_B d'items qui sont vrais en ce questionnaire B et les quatre autres items ne sont pas des items clés. Sauf si une réponse au questionnaire B est $C_B = 0$ ou $C_B = 5$, les
- $$\hat{\pi}_{ps} = \sum_{G+1}^{\infty} c \frac{n_B}{n_B} - \sum_G^{\infty} c \frac{n_A}{n_A} = \sum_{G+1}^{\infty} c \frac{n_B}{n_B} - \sum_G^{\infty} c \frac{n_A}{n_A} \tag{4}$$

Estimateurs de domaine pour la technique du dénombrement d'items

Takahiro Tsuchiya¹

Résumé

La technique du dénombrement d'items (*item count technique*) est une méthode de questionnement indirect qui a été conçue pour estimer la proportion de personnes pour lesquelles un item important de nature délicate est vrai. Elle consiste à demander aux répondants d'indiquer, parmi une liste de phrases descriptives, le nombre d'entre elles qu'ils estiment s'appliquer à eux. Une liste comprenant l'item clé est présentée à une moitié de l'échantillon et une liste ne le contenant pas est présentée à l'autre moitié. La différence entre les nombres moyens de phrases sélectionnées est un estimateur de la proportion recherchée. Dans le présent article, nous proposons deux nouvelles méthodes, appelées méthode par croisement et méthode par double croisement, où les proportions dans les sous-groupes ou domaines sont estimées d'après les données obtenues par la technique du dénombrement d'items. Afin d'évaluer la précision des méthodes proposées, nous réalisons des expériences par simulation au moyen de données tirées d'une enquête sur le caractère national japonais. Les résultats montrent que la méthode par double croisement est beaucoup plus précise que la méthode stratifiée traditionnelle et moins susceptible de produire des estimations illogiques.

Mots clés : Techniques de questionnement indirect; technique du dénombrement d'items; estimateurs de domaine; enquête sur le caractère national japonais.

1. Introduction

1.1 Techniques de questionnement indirect

Supposons qu'une population U soit divisée en deux sous-populations $U^{(T)}$ et $U^{(C)}$, où $U^{(T)}$ est un ensemble d'éléments ayant un attribut T , et $U^{(C)}$ est un complément de $U^{(T)}$. L'un des objectifs des enquêtes sociales est d'estimer $\pi = \frac{Y}{N} = P(Y=1)$, où

$$Y_k = \begin{cases} 1 & \text{si } k \in U^{(T)} \\ 0 & \text{autrement} \end{cases}$$

et $P(\cdot)$ représente la proportion d'unités ayant une valeur particulière de la variable. Par exemple, quand T représente « appuie le cabinet actuel », π indique le taux de soutien pour le cabinet et quand T signifie « consomme une drogue illicite particulière », π représente le taux de prévalence de la consommation de cette drogue.

Dans le cas d'une technique de questionnement direct, les chercheurs demandent aux répondants « Appartenez-vous à $U^{(T)}$? » et obtiennent directement la valeur indicatrice y_i sous forme d'une réponse « Oui » ou « Non » (Cochran 1977, page 50). Si la probabilité de sélection est la même pour tous les répondants, la moyenne d'échantillon \bar{y} sert d'estimateur de π .

Par ailleurs, certaines techniques de questionnement indirect, y compris la technique de la réponse aléatoire (Warner 1965), la technique nominative (Miller 1985), la technique du dénombrement d'items (Droicour, Caspar, Hubbard, Parsley, Visser et Ezzati 1991) et la technique

des trois cartes (Droicour, Larson et Scheuren 2001), sont conçues pour contourner le fait que certains répondants essayent d'éviter les questions délicates, comme celles portant sur des sujets très intimes, des comportements socialement inacceptables ou pervers, ou des actes illégaux. La caractéristique essentielle des techniques indirectes est qu'au lieu d'observer directement X_i , on observe une autre variable $X = g(Y, V)$, qui est une certaine fonction de Y et, au besoin, d'autres variables aléatoires V , de sorte que les répondants aient l'impression que leur réponse réelle pour Y n'est pas révélée. Bien que cette caractéristique permette, en principe, de dériver une réponse correcte pour des répondants évasifs, les procédures tant de questionnement que d'estimation sont assez compliquées comparativement à la technique de questionnement direct, en partie parce que la fonction $g(\cdot)$ comprend parfois des processus de randomisation. Nous décrivons deux techniques indirectes dans les grandes lignes plus loin.

La technique de la réponse aléatoire est la plus populaire des techniques indirectes et diverses versions ont été proposées (Abul-Elaj, Greenberg et Horvitz 1967; Warner 1971; Chaudhuri et Mukerjee 1988; Greenberg, Abul-Elaj, Simmons et Horvitz 1969; Takahasi et Sakasegawa 1977). Bien que cette technique ne soit pas le sujet du présent article, nous décrivons brièvement la procédure originale de Warner à titre de référence, car nous en décrivons la simulation à la dernière section.

1. Préparer deux types de questionnaires. Le questionnaire A comprend la question « Appartenez-vous à

Remerciements

Les auteurs remercient le rédacteur associé, les examinateurs, Paul Smith et Rachel Vis-Vissechers de leurs commentaires constructifs au sujet d'ébauches antérieures du présent article. Jan remercie aussi les professeurs Stephen E. Fienberg et Peter Koopman de leur soutien à titre de conseillers de thèse de doctorat durant les présents travaux.

Bibliographie

- Bethlehem, J.G., et Keller, W.G. (1987). Linear weighing of sample survey data. *Journal of Official Statistics*, 3(2), 141-153.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Fellgett, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Fienberg, S.E., et Tanur, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Revue Internationale de Statistique*, 55(1), 75-96.
- Fienberg, S.E., et Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *The Canadian Journal of Statistics*, 16(2), 135-151.
- Fienberg, S.E., et Tanur, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 243, 1017-1022.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361-374.
- Hájek, J. (1971). Comment on a paper by D. Basu. Dans *Foundations of Statistical Inference*. (Éds. V.P. Godambe et D.A. Sprott). Toronto: Holt, Rinehart et Winston, 236.
- Hartley, H.O., et Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. Dans *Survey Sampling and Measurement*. (Éds. N.K. Namboodiri). New York: Academic Press, 35-43.
- Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York: McGraw-Hill.
- Montgomery, D.C. (2001). *Design and Analysis of Experiments*. New York: John Wiley & Sons, Inc.
- Montesson, D.F. (1990). *Multivariate Statistical Methods*. Singapore: McGraw-Hill.
- Sämdal, C.-E., Swensson, B. et Wierman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons, Inc.
- Seale, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.
- Skinner, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*. (Éds. C.J. Skinner, D. Holt et T.M.F. Smith). Chichester: Wiley & Sons, Inc. 59-87.
- Statcorp. (2001). *Stata Reference Manual Release 7.0*. College Station, Texas.
- Van den Brakel, J.A. (2001). Design and Analysis of Experiments Embedded in Complex Sample Surveys. Thèse de doctorat. Rotterdam: Erasmus University of Rotterdam.
- Van den Brakel, J.A. et Binder, D. (2000). Variance estimation for experiments embedded in complex sampling schemes. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Indianapolis, August 13-17, 805-810.
- Van den Brakel, J.A., et Binder, D. (2004). Variance estimation for experiments embedded in complex sampling designs. Article de recherche non publié, BPA nr.: H894-04-TMO. Heerlen: Statistics Netherlands.
- Van den Brakel, J.A. (1998). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14(3), 277-295.
- Van den Brakel, J.A., et Van Berkel, C.A.M. (2002). A design-based analysis procedure for two-treatment experiments embedded in sample surveys. An application in the Dutch labor force survey. *Journal of Official Statistics*, 18(2), 217-231.

$$\hat{E}_{k;HT} = \sum_{i=1}^{n_{++}} \left(\mathbf{p}_{ik}' (\mathbf{Y}_i - \mathbf{B}' \mathbf{x}_i) \right) \left(\frac{\pi_i N}{\sum_{i=1}^{n_{++}} \mathbf{p}_{ik}' \mathbf{e}_i} \right). \quad (48)$$

En nous servant de (43) et (46), nous pouvons élaborer les éléments diagonaux de $\text{Cov}_e(\hat{\mathbf{E}}_{HT} | m, s)$ comme suit

$$\text{Var}_e(\hat{E}_{k;HT} | m, s) = \text{Cov}_e \left(\sum_{i=1}^{n_{++}} \mathbf{p}_{ik}' \mathbf{e}_i, \sum_{i'=1}^{i'=n_{++}} \mathbf{p}_{i'k}' \mathbf{e}_{i'} | m, s \right)$$

$$= \sum_{j=1}^J \sum_{i'=1}^{i'=n_{++}} \left(\mathbf{e}_i' \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k} | m, s) \frac{\pi_i N}{\mathbf{e}_i'} + \sum_{i=1}^{i=n_{++}} \sum_{i' \neq i}^{i' \neq i} \mathbf{e}_i' \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'k} | m, s) \frac{\pi_i N}{\mathbf{e}_i'} \right)$$

$$= \sum_{j=1}^J \left(\frac{n_{j+} - 1}{n_{j+}} \frac{n_{jk}}{n_{++}} \sum_{i=1}^{i=n_{++}} \left(\frac{\pi_i N}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} \pi_i N \right) \right. \\ \left. - \frac{n_{j+} - 1}{n_{j+}} \sum_{i=1}^{i=n_{++}} \left(\frac{\pi_i N}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} \pi_i N \right) \right) \quad (49)$$

En nous servant de (44) et (45), nous pouvons élaborer les éléments hors diagonale de $\text{Cov}_e(\hat{\mathbf{E}}_{HT} | m, s)$ comme suit

$$\text{Cov}_e(\hat{E}_{k;HT}, \hat{E}_{l;HT} | m, s)$$

$$= \text{Cov}_e \left(\sum_{i=1}^{n_{++}} \mathbf{p}_{ik}' \mathbf{e}_i, \sum_{i'=1}^{i'=n_{++}} \mathbf{p}_{i'l}' \mathbf{e}_{i'} | m, s \right)$$

$$= \sum_{j=1}^J \sum_{i=1}^{i=n_{++}} \left(\mathbf{e}_i' \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'l} | m, s) \frac{\pi_i N}{\mathbf{e}_i'} + \sum_{i' \neq i}^{i' \neq i} \sum_{i'=1}^{i'=n_{++}} \mathbf{e}_i' \text{Cov}_e(\mathbf{p}_{ik}, \mathbf{p}_{i'l} | m, s) \frac{\pi_i N}{\mathbf{e}_i'} \right)$$

$$= \sum_{j=1}^J \left(\frac{n_{j+} - 1}{n_{j+}} \frac{n_{jl}}{n_{++}} \sum_{i=1}^{i=n_{++}} \left(\frac{\pi_i N}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} \pi_i N \right) \right.$$

$$\left. \left(\frac{\pi_i N}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} \pi_i N \right) \right)$$

(50)

Les résultats (49) et (50) peuvent s'écrire en notation matricielle:

$$\bar{\Delta}_f = \sum_{i=1}^J \frac{N}{\mathbf{e}_i}$$

La partie finale de la preuve consiste à prendre l'espérance de $\text{Cov}_e(\hat{\mathbf{C}}_{HT}^e | m, s)$ par rapport au plan d'échantillonnage et au modèle de l'erreur de mesure. La preuve est donnée pour les PBR où les UPE sont les variables de bloc. Selon un plan d'échantillonnage à deux degrés, nous tirons J blocs ou UPE d'une population finie de J_u blocs avec probabilités d'inclusion de premier ordre π_i^j . Dans chaque UPE, nous tirons n_{j+} USE au deuxième degré avec probabilités d'inclusion de premier et de deuxième ordres π_{ij}^{jj} et π_{ij}^{jj} . Les probabilités d'inclusion de premier ordre des individus dans l'échantillon sont $\pi_i = \pi_i^j \pi_{ij}^{jj}$. En outre, soit

$$\bar{\Delta}_f = \sum_{i=1}^J \frac{N}{\mathbf{e}_i}$$

la moyenne de population des erreurs de mesure des individus du bloc j . Alors

$$\bar{\Delta}_f = \sum_{i=1}^J \frac{N \pi_{ij}^{jj}}{\mathbf{e}_i}$$

est l'estimateur d'Horvitz-Thompson pour $\bar{\Delta}_f$. Maintenant, nous avons

$$\text{Cov}_e(\hat{\mathbf{C}}_{HT}^e | m, s) = \mathbf{C} \mathbf{D} \mathbf{C}' - \sum_{j=1}^J \frac{n_{j+}}{n_{++}} - 1 \sum_{i=1}^{i=n_{++}} \left(\mathbf{C} \mathbf{e}_i' \right) \left(\frac{N \pi_i}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} N \pi_{i'} \right) \\ - \sum_{i=1}^{i=n_{++}} \left(\mathbf{C} \mathbf{e}_i' \right) \left(\frac{N \pi_i}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} N \pi_{i'} \right). \quad (51)$$

Il découle de (23) que

$$d_k^k = \sum_{j=1}^J \frac{n_{j+}}{n_{++}} - 1 \sum_{i=1}^{i=n_{++}} \left(\mathbf{Y}_{ik} - \mathbf{b}_i' \mathbf{x}_i \right) \left(\frac{N \pi_i}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} N \pi_{i'} \right).$$

où \mathbf{D} représente une matrice diagonale de dimensions $K \times K$ avec éléments

$$\text{Cov}_e(\hat{\mathbf{E}}_{HT}^e | m, s) = \mathbf{D} - \sum_{j=1}^J \frac{n_{j+}}{n_{++}} - 1 \sum_{i=1}^{i=n_{++}} \left(\mathbf{y}_{ik} - \mathbf{B}' \mathbf{x}_i \right) \left(\frac{N \pi_i}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} N \pi_{i'} \right) \\ - \sum_{i=1}^{i=n_{++}} \left(\mathbf{y}_{ik} - \mathbf{B}' \mathbf{x}_i \right) \left(\frac{N \pi_i}{\mathbf{e}_i} - \frac{1}{n_{j+}} \sum_{i'=1}^{i'=n_{++}} N \pi_{i'} \right)$$

particulier, puisqu'un PRT peut être considéré comme un PBR à un bloc. Notons a. pr. pour « avec probabilité ».

$$\mathbf{p}_{jk} \mathbf{p}_{t'k}^{\prime} = \begin{cases} \left(\frac{n_{j+}}{n_{jk}} \right)^2 \mathbf{r}_k \mathbf{r}_{t'}^k & \text{a. pr.: } \frac{n_{jk}}{n_{j+}} \text{ si } t \in s_j \\ \mathbf{O} & \text{a. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \end{cases}$$

$$\mathbf{p}_{jk} \mathbf{p}_{t'k}^{\prime} = \begin{cases} \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \mathbf{r}_k \mathbf{r}_{t'}^k & \text{a. pr.: } 0 \text{ si } t \in s_j \\ \mathbf{O} & \text{a. pr.: } 1 \end{cases}$$

$$\mathbf{p}_{jk} \mathbf{p}_{t'k}^{\prime} = \begin{cases} \frac{n_{j+}}{n_{jk}} \frac{n_{t'+}}{n_{jk}} \mathbf{r}_k \mathbf{r}_{t'}^k & \text{a. pr.: } \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \\ \frac{n_{j+}}{n_{jk}} \frac{n_{t'+}}{n_{jk}} \mathbf{r}_k \mathbf{r}_{t'}^k & \text{a. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \\ \mathbf{O} & \text{a. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \end{cases}$$

$$E^e(\mathbf{p}_{jk}) = P \left(\mathbf{p}_{jk} = \frac{n_{j+}}{n_{jk}} \mathbf{r}_k + P(\mathbf{p}_{jk} = \mathbf{0}) \mathbf{0} = \mathbf{r}_k. \right) \quad (42)$$

Nous pouvons établir les covariances qui suivent par rapport au plan d'expérience :

L'espérance de \mathbf{p}_{jk} par rapport au plan d'expérience est donnée par :

$$\mathbf{p}_{jk} \mathbf{p}_{t'k}^{\prime} = \begin{cases} \left(\frac{n_{j+}}{n_{jk}} \right)^2 \mathbf{r}_k \mathbf{r}_{t'}^k & \text{a. pr.: } \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \\ \mathbf{O} & \text{a. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \\ \frac{n_{j+}}{n_{jk}} \frac{n_{t'+}}{n_{jk}} \mathbf{r}_k \mathbf{r}_{t'}^k & \text{a. pr.: } \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \\ \mathbf{O} & \text{a. pr.: } 1 - \frac{n_{jk}}{n_{j+}} \frac{n_{jk}}{n_{t'+}} \text{ si } t \in s_j, t' \in s_j \end{cases}$$

Preuve de la formule (23)

Sous la condition énoncée qu'il existe un vecteur \mathbf{a} de dimension H constant tel que $\mathbf{a}' \mathbf{x}_i = 1$ pour tout $i \in U$, et sachant la réalisation de $u_i, i = 1, \dots, N$, il découle du modèle de superpopulation (16) que \mathbf{b}_k dans (18) peut être évalué sous la forme

$$E^m(\mathbf{b}_k) = E^m \left(\sum_{i=1}^N \frac{\omega_i^2}{\mathbf{x}_i' \mathbf{y}_{ik}} \right) = E^m \left(\sum_{i=1}^N \frac{\omega_i^2}{\mathbf{x}_i' \mathbf{x}_i'} \right) = \mathbf{b} + \mathbf{d} + \mathbf{a} \beta_k \quad (47)$$

où \mathbf{b} représente les coefficients de régression définis par (17) et \mathbf{d} , les coefficients de régression de la fonction de régression des effets d'intervieweur sur les variables auxiliaires $\mathbf{x}_{t'}$. Du résultat (47) il découle que $\mathbf{B}' \mathbf{x}_i = \mathbf{j}(\mathbf{b}' \mathbf{x}_i + \mathbf{d}' \mathbf{x}_i) + \beta$. Puisque $\mathbf{C} \mathbf{j} = \mathbf{0}$, et d'après le modèle de l'erreur de mesure (1) et le modèle de régression linéaire (16), il découle que

$$\mathbf{C}(\mathbf{y}_i - \mathbf{B}' \mathbf{x}_i) = \mathbf{C}(\mathbf{j} u_i + \mathbf{j} \psi_i + \beta + \varepsilon_i - \mathbf{j}(\mathbf{b}' \mathbf{x}_i + \beta)) = \mathbf{C} \varepsilon_i, \quad \mathbf{C} \mathbf{Q}, \mathbf{D}, \mathbf{F}.$$

Preuve de la formule (26) pour un PBR

Nous commençons par établir une expression pour $\text{Cov}^e(\mathbf{C} \hat{\mathbf{B}}_{\text{HT}} | m, s)$. Soit $\mathbf{e}_i = (e_{i1}, \dots, e_{ik})'$ un vecteur de dimension K avec éléments $e_{ik} = y_{ik} - \mathbf{b}'_k \mathbf{x}_{t'}$. Par conséquent, $\mathbf{e}_i = \mathbf{y}_i - \mathbf{B}'_s \mathbf{x}_{t'}$. Notons que $E^m \mathbf{B}_s \text{Cov}^e(\mathbf{C} \hat{\mathbf{B}}_{\text{HT}} | m, s) = \mathbf{C} E^m \mathbf{B}_s \text{Cov}^e(\hat{\mathbf{B}}_{\text{HT}} | m, s) \mathbf{C}'$ avec $E^m \hat{\mathbf{B}}_{\text{HT}} = (E^m_{1;\text{HT}}, \dots, E^m_{K;\text{HT}})'$. En outre, notons que

d'échantillonnage étaient tirées avec remise et avec probabilités de sélection inégales. Ni les probabilités d'inclusion de deuxième ordre ni les covariances par rapport au plan de sondage entre les estimations sur sous-échantillon ne doivent être connues, ce qui simplifie considérablement l'analyse. Ainsi, dans le cas de l'échantillonnage aléatoire simple avec remise, ce résultat signifie que l'on devrait laisser tomber le facteur de correction pour population finie dans l'estimation de la variance des contrastes. Par conséquent, nous obtenons une statistique de Wald, établie par des ajustements du plan de sondage sous des plans de sondage complexes généraux, qui retiennent la structure assez simple intéressante des méthodes classiques d'analyse fondées sur un modèle.

Pour les PRT et les PBR intégrés dans un plan d'échantillonnage autopondéré analysés au moyen de l'estimateur étendu d'Horvitz-Thompson et d'un estimateur groupé de la variance, la statistique de Wald coïncide avec la statistique F d'une analyse de variance à un ou à deux critères de classification. Pour l'analyse de l'expérience à deux traitements intégrés, on peut établir une version fondée sur le plan de sondage de la statistique t en tant que cas particulier de la statistique de Wald. Les expressions et des renseignements supplémentaires au sujet de cette statistique t fondée sur le plan de sondage et de sa relation avec la statistique de Welch et avec la statistique t standard figurent dans Van den Brakel et Renssen (1998), Van den Brakel (2001) ou Van den Brakel et Van Berkel (2002).

La méthode d'analyse proposée dans le présent article est implémentée dans un logiciel appelé X-tool. Cet outil sera disponible en tant que composante du logiciel de traitement des données d'enquête Blaise développé par Statistique Pays-Bas.

Annexe

Propriétés des vecteurs de randomisation \mathbf{p}_{jk}

Pour les PRT et les PBR, les vecteurs de randomisation \mathbf{p}_{jk} sont définis par (14) et (15). En raison du mécanisme de randomisation du plan d'expérience, les vecteurs \mathbf{p}_{jk} sont aléatoires et ont les fonctions de masse de probabilité conditionnelle qui suivent. Pour un PRT, nous avons

$$P\left(\mathbf{p}_{jk} = \frac{n_{+}^k}{n_{+}} \mathbf{r}_k \mid s\right) = \frac{n_{+}^k}{n_{+}} \text{ et } P(\mathbf{p}_{jk} = 0 \mid s) = 1 - \frac{n_{+}^k}{n_{+}}.$$

Pour un PBR, nous avons

$$P\left(\mathbf{p}_{jk} = \frac{n_{+}^k}{n_{+}} \mathbf{r}_k \mid s_j\right) = \frac{n_{+}^k}{n_{+}} \text{ et } P(\mathbf{p}_{jk} = 0 \mid s_j) = 1 - \frac{n_{+}^k}{n_{+}}.$$

Nous établissons les propriétés de ces vecteurs pour un PBR. Les propriétés pour un PRT en découlent comme cas

de sondage, cette approche produit des valeurs p plus grandes pour le test des effets de traitement.

Un autre avantage important du test de Wald fondé sur le plan de sondage comparativement à l'approche par la régression linéaire fondée sur le plan de sondage est qu'il a toujours trait aux différences entre les estimations sur les sous-échantillons, ce qui facilite l'interprétation des résultats. Cette propriété est particulièrement importante pour les expériences intégrées visant à quantifier des ruptures de tendance concernant les paramètres d'une enquête causée par des ajustements du plan de sondage. Dans le cas d'un plan en randomisation totale, le modèle de régression linéaire est constitué d'une ordonnée à l'origine et de trois coefficients pour les effets de traitement. Dans cette situation particulièrement simple, les coefficients des effets de traitement sont exactement égaux aux différences entre les estimations sur les sous-échantillons. Cependant, cette propriété ne tient pas pour les effets de traitement obtenus sous des modèles plus complexes, comme le cas du plan en blocs randomisés.

5. Discussion et conclusions

Nous discutons dans le présent article de la manière dont la méthodologie statistique des expériences randomisées et de l'échantillonnage aléatoire peut appuyer la conception et l'analyse d'expériences intégrées dans les enquêtes par sondage courantes. Le plan de l'enquête par sondage constitue un cadre *a priori* pour l'application de principes, titres de la théorie des plans expérimentaux, tels que la randomisation et le contrôle local par constitution de blocs sur les strates, les UPE, les grappes ou les interviewers. Pour tester les hypothèses au sujet des estimations des paramètres de population finie obtenues sous divers traitements de l'expérience, nous établissons une statistique de Wald fondée sur le plan de sondage pour l'analyse des plans en randomisation totale et des plans en blocs randomisés intégrés dans des plans d'échantillonnage complexes généraux en utilisant l'estimateur d'Horvitz-Thompson et l'estimateur par la régression généralisée. La combinaison de l'échantillonnage aléatoire d'une population finie et de cette méthode d'analyse fondée sur le plan de sondage nous permet de généraliser les résultats de l'expérience observés dans l'échantillon particulier à l'ensemble de la population d'enquête.

Puisque nous tenons compte de plans de sondage complexes généraux, nous nous attendons à obtenir une expression assez compliquée pour la matrice des covariances des effets de traitement, avec des éléments hors diagonal non nuls. Cependant, l'estimateur établi pour cette matrice des covariances a la même structure que si les unités

Tableau 5.2

Régression fondée sur le plan, PRT

Source	Coefficient	Erreur-type	Statistique de Wald	ddl	Valeur p
Traitement					
Traitement 1	-182,14	177,60	2,907	3	0,4062
Traitement 2	-58,36	175,56			
Traitement 4	66,79	170,46			
Constante	3 596,47	194,75			

Tableau 5.3

Analyse de variance classique, PRT

k	β_k	\bar{y}_k	Contraste		ANOVA		
1	0	8021	k - k'	$\bar{y}_k - \bar{y}_{k'}$	Source	ddl	F
2	80	8094	1 - 2	-73	Entre traitements	3	14 432 816
3	160	7955	1 - 3	66	Résidu	3 776	104 924 668
4	240	8242	1 - 4	-221	Total	3 779	

Tableau 6.1

Statistique de Wald fondée sur le plan, PBR

Sous-échantillons				Contrastes			
k	β_k	\bar{y}_k	k - k'	$\bar{y}_k - \bar{y}_{k'}$	$\sqrt{d_k + d_{k'}}$	W	ddl
1	0	3 395	1 - 2	-25	81,247	9,93011	3
2	80	3 420	1 - 3	-120	80,697	0,0192	
3	160	3 515	1 - 4	-231	82,383		
4	240	3 626					

Tableau 6.2

Régression fondée sur le plan, PBR

Source	Coefficient	Erreur-type	Statistique de Wald	ddl	Valeur p
--------	-------------	-------------	---------------------	-----	----------

Bloc							
Bloc 2	-17 068,28	2 556,46	18,4212	3	0,00036		
Bloc 3	-21 999,39	2 540,98					
Traitement 1	-211,51	74,84					
Traitement 2	-246,78	60,05					
Traitement 3	-97,91	73,39					
Constante	23 589,64	2543,25					

Tableau 6.3

Analyse de variance classique, PBR

k	β_k	\bar{y}_k	Contraste		ANOVA		
1	0	8 815	k - k'	$\bar{y}_k - \bar{y}_{k'}$	Source	ddl	F
2	80	8 150	1 - 2	665	Entre blocs	2	1,6773 E+11
3	160	8 566	1 - 3	249	Entre traitements	3	84 377 227
4	240	8 746	1 - 4	69	Résidu	3 774	42 310 035
Total						3 779	131 089 505

Comme nous le soulignons à la section 3, l'approche par la régression linéaire ne tient pas compte de la variance de plan due à la randomisation des unités d'échantillonnage sur les sous-échantillons selon le plan d'expérience. Par conséquent, les erreurs-types des effets de traitement sont plus faibles sous l'approche par la régression linéaire que dans le cas du test de Wald fondé sur le plan, et l'approche de la régression fondée sur le plan produit des valeurs p plus faibles pour le test des effets de traitement.

L'analyse de variance classique est une approche naïve, puisqu'elle ne tient pas compte de la stratification, de la mise en grappes et de la sélection des unités d'échantillonnage en utilisant des probabilités d'inclusion proportionnelles à la valeur du paramètre cible. Omettre de tenir compte de ces aspects du plan d'échantillonnage dans l'analyse a pour résultat net d'exagérer fortement les estimations sur les sous-échantillons, ainsi que les erreurs-types. Comparativement aux deux autres méthodes fondées sur le plan

4.2 Comparaison de trois méthodes d'analyse

De surcroît, nous comparons trois méthodes d'analyse possibles pour les expériences intégrées, c'est-à-dire le test de Wald fondé sur le plan proposé à la section 2, une analyse de variance Anova classique où les observations sont équilibrées et considérées comme étant i.i.d., et l'approche de la régression linéaire fondée sur le plan décrite à la section 3. À cette fin, nous tirons deux échantillons, chacun de taille égale à 3 780 USE, de la population finie spécifiée au tableau 1, selon le plan d'échantillonnage stratifié à deux degrés qui a été utilisé pour la simulation précédente (voir le tableau 2). Nous répartissons les USE de l'un de ces échantillons aléatoirement en quatre sous-échantillons, chacun de taille égale à 945, selon un PRT et les USE de l'autre échantillon aléatoirement en quatre sous-échantillons, chacun de taille égale à 945, selon un PRT et nous tablerons les résultats de l'analyse sous un PBR sont résumés aux tableaux 6.1, 6.2 et 6.3.

Tableau 4.6
Résultats de la simulation PBR $\beta = (0, 20, 40, 60)^T$

Sous-échantillons				Contrastes			
k	β_k	I_k	d_k	$k - k'$	CY	CVC'	CDC' $\sigma(CDC')$
1	0	3 390	3 090	1 - 2	-20	6 225	6 180
2	20	3 410	3 089	1 - 3	-40	6 177	6 181
3	40	3 430	3 090	1 - 4	-60	6 184	6 180
4	60	3 450	3 090				
Éléments diagonaux de				Contrastes			
α	$P(W)$	α	$P(W)$	α	$P(W)$	α	$P(W)$
$p_{sm}(W)$	0,09371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0,050	0,025	0,05096	0,010	0,02365
	0,05238	0,05096	0,010	0,02365	0,05096	0,010	0,02365
	0,02405	0,02365	0,010	0,02365	0,05096	0,010	0,02365
	0,009371	0,09099	0				

statistique de Wald approximativement bien la puissance réelle. En moyenne, la puissance simulée est légèrement plus élevée. L'espérance de la loi du chi-carré est égale à $E(\chi^2_{[K-1]}) = (K-1) + 2\delta$ (Searle 1971, section 2.4h). Si la distribution de rééchantillonnage de la statistique de Wald tend vers une loi $\chi^2_{[K-1] + 2\delta}$, alors la moyenne de la statistique de Wald de rééchantillonnage \bar{W} (40) doit tendre vers l'espérance de la loi du chi-carré. En effet, il découle des tableaux 4.1 à 4.8 que $\bar{W} \approx (K-1) + 2\delta$. De plus, nous vérifions l'hypothèse que la distribution de rééchantillonnage de la statistique de Wald sous l'hypothèse nulle est égale à la loi du chi-carré centrée situation particulière.

Tableau 4.1
Résultats de la simulation PRT $\beta = (0, 0, 0, 0)^T$

Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	0	3 392	14 311	1-2	0	28 725	28 616
3	0	3 392	14 306	1-3	0	28 892	28 616
2	0	3 392	14 306	1-4	2	28 787	28 603
1	0	3 392	14 311				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	0	3 390	14 292	1-4	2	28 787	28 603
3	0	3 392	14 306	1-3	0	28 892	28 616
2	0	3 392	14 306	1-4	2	28 787	28 603
1	0	3 392	14 311				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	0	3 390	14 292	1-4	2	28 787	28 603
3	0	3 392	14 306	1-3	0	28 892	28 616
2	0	3 392	14 306	1-4	2	28 787	28 603
1	0	3 392	14 311				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	0	3 392	14 307	1-2	-20	28 635	28 614
3	40	3 432	14 314	1-3	-40	28 918	28 620
2	20	3 412	14 307	1-4	-58	28 624	28 597
1	0	3 392	14 307				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	60	3 450	14 291	1-4	-58	28 624	28 597
3	40	3 432	14 314	1-3	-40	28 918	28 620
2	20	3 412	14 307	1-4	-58	28 624	28 597
1	0	3 392	14 307				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	120	3 511	14 295	1-4	-119	28 713	28 609
3	80	3 472	14 307	1-3	-80	28 947	28 622
2	40	3 432	14 307	1-2	-40	28 597	28 621
1	0	3 392	14 314				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616
1	0	3 392	14 306				
Sous-échantillons				Contrastes			
k	β_k	\bar{Y}_k	d_k	$k-k'$	CY	CVC'	$\sigma(CBC')$
4	240	3 631	14 291	1-4	-239	28 538	28 598
3	160	3 552	14 312	1-3	-160	28 784	28 618
2	80	3 472	14 310	1-2	-80	28 748	28 616</

Une approximation de la matrice réelle des covariances des effets de traitement est donnée par

$$CVC' = \frac{1}{R-1} \sum_{r=1}^R C(\tilde{Y}^r - \bar{Y})(\tilde{Y}^r - \bar{Y})'C' \quad (41)$$

Nous évaluons la performance de la méthode d'estimation de la variance par comparaison de CBC' à CVC' . Si l'estimateur de la variance établi CBC' est approximativement sans biais par rapport au plan, alors la moyenne des matrices des covariances de rééchantillonnage CBC' doit tendre vers la matrice réelle des covariances CVC' , pour $R \rightarrow \infty$. Le calcul de l'écart-type des éléments de CBC' , noté $\sigma(CBC')$, donne une idée de la précision de l'estimateur de la variance établi. Les éléments diagonaux de D sont notés d_k .

Si $\tilde{Y}^{CVC'} \rightarrow N(C\beta, CVC')$, il s'ensuit que

$$W \rightarrow \chi^2_{[K-1]|\delta|}, \text{ où } K-1 \text{ est le nombre de degrés de liberté et } \delta = 1/2(C\beta)'(CVC')^{-1}(C\beta)$$

non-centralisé de la loi du chi-carré. Dans l'étude en simulation, nous pouvons calculer le paramètre de non-centraité sous les hypothèses alternatives en insérant (41) dans l'expression de δ . Ensuite, nous pouvons calculer la puissance de la statistique de Wald pour un ensemble particulier d'effets de traitement par $P(W) = P(\chi^2_{[K-1]|\delta|} > \chi^2_{[1-\alpha]|\delta|})$, où $\chi^2_{[1-\alpha]|\delta|}$ est le $(1-\alpha)^\circ$ percentile de la loi du chi-carré centrée à $K-1$ degrés de liberté. Nous évaluons les propriétés de la statistique de Wald en comparant $P(W)$ à la puissance simulée, qui est définie comme étant la fraction d'exécutions significatives observées dans les R rééchantillonnages, c'est-à-dire

$$P^{sim}(W) = \frac{1}{R} \sum_{r=1}^R I(W^r > \chi^2_{[1-\alpha]|\delta|}),$$

où $I(B)$ est la variable indicatrice qui est égale à l'unité si B est vrai et nulle autrement. Les résultats des simulations sont résumés aux tableaux 4.1 à 4.8.

Les moyennes des estimations sur les sous-échantillons \tilde{Y}_k sous l'hypothèse nulle présentées aux tableaux 4.1 et 4.5 surestiment légèrement la moyenne de population donnée au tableau 1. Cette différence peut être attribuée au biais de l'estimateur étendu d'Horvitz-Thompson. Par contre, les moyennes des contrastes entre les estimations sur les sous-échantillons $C\tilde{Y}$ concordent presque parfaitement avec les effets de traitement réels $C\beta$. Les moyennes des matrices des covariances de rééchantillonnage CBC' tendent vers les valeurs des matrices des covariances réelles CVC' , résultat qui montre que la méthode d'estimation de la variance établie à la section 2.4 est approximativement sans biais par rapport au plan. La précision relative des éléments diagonaux de CBC' est d'environ 10,5 % sous cette taille d'échantillon particulière. La puissance simulée fondée sur la distribution de rééchantillonnage de la

hypothèses alternatives distinctes. On obtient ainsi huit simulations différentes, dont les spécifications sont données au tableau 3. Chaque simulation est fondée sur $R = 100\,000$ rééchantillonnages. Les observations du paramètre cible sont obtenues en ajoutant une erreur de mesure et un effet de traitement aux valeurs intrinsèques conformément à (39).

Tableau 3

Sommaire des conditions de simulation

Effets de traitement		Plan d'expérience			
PRT	PBR	0	0	β_1	β_4
PRT	PBR	0	20	β_2	
PRT	PBR	0	40	β_3	
PRT	PBR	0	80		
PRT	PBR	0	160		
PRT	PBR	0	240		

Les données obtenues lors de chaque rééchantillonnage sont analysées au moyen de l'estimateur d'Horvitz-Thompson étendu (30). Soit \tilde{Y}_k^r l'estimation sur sous-échantillon obtenue sous le k° traitement dans le r° rééchantillonnage. Le vecteur contenant les quatre estimations sur sous-échantillon obtenues lors du r° rééchantillonnage est donné par $\tilde{Y}^r = (\tilde{Y}_1^r, \tilde{Y}_2^r, \tilde{Y}_3^r, \tilde{Y}_4^r)'$. Le vecteur avec les trois contrastes pour le r° rééchantillonnage est égal à $C\tilde{Y}^r$, avec $C = (f; -1, 1)$ un vecteur de dimension 3 dont chaque élément est égal à 1 et I la matrice identité de dimensions 3×3 . En outre, d_k^r représente les éléments diagonaux de la matrice des covariances estimée, obtenue sous le r° rééchantillonnage. Une expression de d_k^r est donnée par (29) avec $b_k^r x_i = \tilde{Y}_i^r$. La matrice estimée des covariances des effets de traitement est égale à CBC' , avec $D' = \text{diag}(d_1^r, d_2^r, d_3^r, d_4^r)$. Enfin, $W^r = (C\tilde{Y}^r)'(CBC')^{-1}(C\tilde{Y}^r)$ représente la statistique de Wald observée lors du r° rééchantillonnage. D'après les $R = 100\,000$ rééchantillonnages pour chaque simulation, les paramètres de population sous les divers traitements peuvent être approximés par

$$\bar{Y} = \frac{1}{R} \sum_{r=1}^R \tilde{Y}^r,$$

avec $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4)'$. De (10) il découle que les effets de traitement réels dans le modèle de mesure peuvent être approximés par $C\bar{Y} \approx C\beta$. En outre, la moyenne des matrices des covariances de rééchantillonnage peut être calculée selon

$$CBC' = \frac{1}{R} \sum_{r=1}^R CBC',$$

et la moyenne des statistiques de Wald de rééchantillonnage selon

$$\bar{W} = \frac{1}{R} \sum_{r=1}^R W^r. \quad (40)$$

étape. Dans chaque strate nous appliquons des bornes supérieures et inférieures et des larges d'intervalle différentes pour ces lois uniformes, de sorte que la population puisse être stratifiée en trois sous-populations relativement homogènes. Les intervalles des lois uniformes qui sont appliquées à la deuxième étape sont plus petits que les intervalles des lois uniformes appliquées à la première étape. Il en résulte une population où les valeurs intrinsèques pour les USE contenues dans chaque UPE sont mises en grappes. La structure de la population est résumée au tableau 1.

Tableau 1
Population

Valeur intrinsèque du paramètre cible		Population	
Strate	Nombre	Nombre	Valeur
1	70	6 250	min.
2	130	18 250	max.
3	250	6 128	max.
Total	450	109 500	max.

Nous tirons des échantillons de façon répétée de cette population selon un plan d'échantillonnage stratifié à deux degrés sans remise avec probabilités d'inclusion inégales. Les probabilités d'inclusion sont choisies proportionnellement à la taille du paramètre cible. Les tailles d'échantillon pour les diverses strates sont résumées au tableau 2. Pour chaque échantillon, nous générons une nouvelle erreur de mesure pour chaque élément de population. Ces erreurs de mesure sont tirées d'une loi normale de moyenne nulle et d'écart-type proportionnel à la grandeur des valeurs intrinsèques. La fourchette des écarts-types varie de 1 000 pour les USE avec les valeurs intrinsèques les plus grandes dans la première strate à 10 pour les USE avec les valeurs intrinsèques les plus petites dans la troisième strate.

Tableau 2
Plan d'échantillonnage

Strate	Nombre d'UPE	Nombre d'USE
1	25	900
2	30	1 080
3	50	1 800
Total	105	3 780

Enfin, nous subdivisons aléatoirement les échantillons conformément à un plan d'expérience, en quatre sous-échantillons contenant chacun 945 USE. Nous appliquons deux plans d'expérience distincts. Dans le premier, les USE sont randomisées sur les quatre traitements différents selon un PRT. Dans le deuxième, elles sont randomisées sur les quatre traitements différents selon un PBR, où les trois strates sont utilisées comme variables de bloc. Dans chaque bloc ou strate, un quart des USE est assigné aléatoirement à chaque traitement. Sous les deux plans d'expérience, quatre ensembles distincts d'effets de traitement sont appliqués, un sous l'hypothèse nulle et les trois autres sous des

$$\lambda^{qj} = \sum_{i \in U} \sum_{t \in U} (\pi_i^t - \pi_i \pi_t) \frac{x_i^{tq} x_t^{jq}}{x_i^{tq} x_t^{jq} \pi_i^t \pi_t^j} + \sum_{i \in U} \sum_{t \in U} \pi_i^t \pi_t^j (x_i^{tq} x_t^{jq} - \pi_i^t \pi_t^j) \frac{x_i^{tq} x_t^{jq}}{x_i^{tq} x_t^{jq} \pi_i^t \pi_t^j}$$

qui a la structure de variance d'un échantillon à deux phases, où la première phase correspond au plan d'échantillonnage et la deuxième, au plan d'expérience. Les unités d'échantillonnage sont, suivant le plan d'expérience, assignées à un seul des K traitements. Il s'ensuit que $\pi_i^{tj} = 0$ pour $k \neq k'$, et $i = i'$, ce qui rend difficile l'établissement d'un estimateur approximativement sans biais par rapport au plan pour les termes de covariance de $\text{Var}(\beta)$; voir aussi Van den Brakel et Binder (2000, 2004). Dans la méthode d'analyse proposée à la section 2, ce problème est contourné en établissant un estimateur fondé sur le plan pour la matrice des covariances des contrastes de C^{REG} au lieu d'un estimateur de la matrice des covariances de Y^{REG} proprement dit.

4. Étude en simulation

À la sous-section 4.1, nous procédons à une étude en simulation en vue d'évaluer la performance de l'estimateur fondé sur le plan de la matrice des covariances des contrastes entre les estimations sur sous-échantillon CDC' avec éléments diagonaux donnés par (29), ainsi que la statistique de Wald fondée sur le plan W définie par (32) pour tester les hypothèses au sujet de ces contrastes. Puis, à la sous-section 4.2, nous appliquons ce test de Wald fondé sur le plan, l'approche de la régression linéaire fondée sur le plan et une analyse de variance classique Anova à l'analyse d'un plan en randomisation totale et d'un plan en blocs randomisés.

4.1 Évaluation de l'absence de biais dans CDC' et de la loi de W

Dans la présente étude en simulation, nous supposons que le modèle de l'erreur de mesure ne contient pas d'effet d'intervieweur, c'est-à-dire que

$$y_{ik} = u_i + \beta_k + \varepsilon_{ik} \quad (39)$$

Nous générons une population artificielle constituée de trois strates, 450 UPE et 109 500 USE par tirage aléatoire de valeurs strictement positives pour les valeurs intrinsèques u_i d'un paramètre cible. Les tailles des UPE dans la population sont inégales. Nous générons les valeurs intrinsèques en deux étapes. Premièrement, nous tirons une valeur positive pour chaque UPE dans la population à partir d'une loi uniforme. Puis, nous tirons une valeur positive pour chaque UPE à la première

3. Analyse par régression linéaire basée sur le

plan de sondage

Nous pourrions envisager une régression linéaire basée sur le plan de sondage pour remplacer l'analyse des expériences intégrées. On suppose, dans ce cas, que les observations sont le résultat d'un modèle de régression linéaire $y_i = B'x_i + e_i$, avec x_i le vecteur contenant Q variables explicatives, B le vecteur contenant les coefficients de régression et e_i un résidu. Ce modèle est déterminé principalement par le plan d'expérience et contient les facteurs de traitement, les facteurs de contrôle locaux (par exemple, blocs) et les covariables comme variables explicatives (voir, par exemple, Montgomery 2001). Des covariables possibles sont les variables auxiliaires du schéma de pondération de l'estimateur par la régression généralisée. Les paramètres d'intérêt sont les coefficients de régression dans la population finie, qui sont définis par $\beta = (X'X)^{-1}X'y$, où X est la matrice de plan d'expérience de dimensions $N \times Q$ et y est un vecteur de dimension N contenant les observations obtenues sous les divers traitements, comme si la population finie complète était incluse dans l'expérience. La matrice de plan d'expérience subdivise conceptuellement la population en K sous-populations ou domaines, qui sont observés sous chacun des K traitements de l'expérience. La taille de chaque sous-population est déterminée par la fraction d'unités d'échantillonnage assignées à chaque traitement dans l'expérience. Un estimateur des coefficients de régression fondé sur le plan est donné par $\hat{\beta} = (X_n'X_n)^{-1}X_n'y_n$, (Särndal et coll. 1992, section 5.10). Ici, X_n est la matrice de plan d'expérience de dimensions $n \times Q$, y_n est un vecteur contenant n observations obtenues sous les divers traitements des n unités incluses dans l'échantillon, et Π est une matrice diagonale de dimensions $n \times n$ contenant les probabilités d'inclusion de premier ordre π_i du plan de sondage. La matrice des covariables approximative de $\hat{\beta}$ est donnée par (Särndal et coll. 1992, section 5.10)

$$\text{Var}(\hat{\beta}) = (X'X)^{-1}V(X'X)^{-1}, \quad (38)$$

avec $V = \text{Var}_s(X_n'\Pi^{-1}y_n - X_n'\Pi^{-1}X_n\beta)$. Les éléments de V sont donnés par

$$\lambda^{pq} = \sum_{i \in U} \sum_{j \in U} (\pi_i \pi_j - \pi_i \pi_j \pi_{ij}) \frac{x_i^p x_j^q}{x_i^p x_j^q e_i e_j}, \quad p, q = 1, \dots, Q,$$

avec $e_i = y_i - \beta'x_i$. Les hypothèses au sujet du sous-ensemble de coefficients de régression qui reflète les effets de traitement sont soumises à un test de Wald; voir, par exemple, Skinner (1989). Le principal inconvénient de cette approche est que la méthode d'estimation ne tient pas compte de l'affectation

aléatoire des unités d'échantillonnage aux traitements conformément aux plans d'expérience. En procédant comme cela, les estimations sur les sous-échantillons sont traitées incorrectement comme s'il s'agissait d'estimations de domaine, ce qui donne des variances par rapport aux plans de sondage incorrectes. La matrice des covariances des effets de traitement (28), établie à la section 2.4, illustre le fait que la superposition du plan d'expérience au plan d'échantillonnage détermine quelles caractéristiques particulières du plan d'échantillonnage sont annulées ou préservées. Par exemple, l'effet de l'échantillonnage stratifié ou de l'échantillonnage à deux degrés sur la variance des effets de traitement est annulé sous un plan en randomisation totale. Toutefois, l'approche de la régression linéaire ignore cet effet, puisque $\text{Var}(\hat{\beta})$ tient compte uniquement de la variance du plan de sondage. Le fait qu'il n'est pas tenu compte du plan d'expérience dans la méthode d'estimation de la variance devient encore plus évident sous un dénombrement complet de la population finie. En raison du plan d'expérience, la population finie entière est subdivisée aléatoirement en K sous-échantillons et les paramètres sous les divers traitements sont encore estimés avec une variance de plan non nulle. Dans cette situation, il s'ensuit, pour l'approche de la régression linéaire, que $\hat{\beta} = \beta$ et que $\text{Var}(\hat{\beta})$ est nulle parce que la variance de plan induite par le plan d'expérience n'est pas prise en compte. Ceci fait contraste avec (28) qui, sous un recensement complet, reflète encore la variance de plan due au plan d'expérience. La façon de corriger l'approche de la régression linéaire pour tenir compte de la randomisation due au plan d'échantillonnage ainsi qu'au plan d'expérience n'est pas évidente. Sachant la réalisation de l'échantillon, le plan d'expérience peut être décrit par des probabilités d'inclusion de premier et de deuxième ordres. Soit π_k^i la probabilité d'inclusion de premier ordre que la i^{e} unité d'échantillonnage soit assignée au k^{e} traitement et soit π_{kl}^{ij} la probabilité d'inclusion de deuxième ordre que la i^{e} unité d'échantillonnage soit assignée au k^{e} traitement et que la j^{e} unité d'échantillonnage soit assignée au l^{e} traitement. Un est-matériau fondé sur le plan de β , qui tient compte du plan de sondage et du plan d'expérience, est donné par $\hat{\beta} = (X_n'\Pi^{-1}X_n)^{-1}X_n'\Pi^{-1}y_n$, où Π^* est la matrice diagonale de dimensions $n \times n$ avec les probabilités d'inclusion de premier ordre $\pi_i^* = \pi_i \pi_k^i$. Une approximation de la matrice des covariables de $\hat{\beta}$ est donnée par (38), où V est obtenue en imposant comme condition la réalisation de l'échantillon, c'est-à-dire

$$V = \text{Var}_s E_e(X_n'\Pi^{-1}y_n - X_n'\Pi^{-1}X_n\beta) + E_s \text{Var}_e(X_n'\Pi^{-1}y_n - X_n'\Pi^{-1}X_n\beta).$$

Ceci mène pour les éléments de V à l'expression :

$$S_{E, \hat{p}_2}^2 = \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{n_j^+ - K} (n_j^+ - K)$$

$$\left(\frac{n_j^+ (Y_{jk} - \mathbf{b}_k^T \mathbf{x}_j)}{N \pi_j} - \frac{1}{n_j^+} \sum_{k=1}^{n_j^+} \frac{N \pi_j}{n_j^+ (Y_{jk} - \mathbf{b}_k^T \mathbf{x}_j)} \right)^2 \quad (34)$$

Dans plusieurs cas particuliers, la statistique de Wald basée sur le plan de sondage coïncide avec la statistique F utilisée dans les méthodes d'analyse basées sur un modèle plus conventionnel. Considérons un PBR intégré dans un plan de sondage autopondéré, où les unités d'échantillonnage sont réparties proportionnellement entre les traitements sur les blocs, c'est-à-dire $\pi_j = n_j^+ / N$ et $n_j^+ / n_j^+ = n_j^+ / n_j^+$ pour tout $j = 1, \dots, J$. Alors, il découle des résultats obtenus pour l'estimateur par le ratio (30) que $\frac{1}{n_j^+} \sum_{k=1}^{n_j^+} Y_{jk} \equiv \bar{Y}_{jk}$ et $\mathbf{b}_k^T \mathbf{x}_j \equiv \bar{Y}_{jk}$. Notons $\bar{Y}_{j+} = 1/n_j^+ \sum_{k=1}^{n_j^+} Y_{jk}$ et $\bar{Y}_{++} = 1/n_j^+ \sum_{k=1}^{n_j^+} \sum_{j=1}^J Y_{jk}$, alors nous

$$\frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{n_j^+} Y_{jk} = \bar{Y}_{j+}, \quad \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{n_j^+} \mathbf{b}_k^T \mathbf{x}_j = \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{n_j^+} \frac{n_j^+}{n_j^+} \bar{Y}_{jk} = \bar{Y}_{++}$$

Si $n_j^+ \approx n_j^+ - 1$, alors il s'ensuit sous l'estimateur groupé de la variance (33) que

$$\hat{d}_k = \sum_{j=1}^J \frac{n_j^+}{n_j^+} \frac{n_j^+}{n_j^+} - 1 \sum_{k=1}^K \sum_{j=1}^{n_j^+} \frac{n_j^+}{n_j^+}$$

$$\left(\frac{Y_{jk} - \mathbf{b}_k^T \mathbf{x}_j}{N \pi_j} - \frac{1}{n_j^+} \sum_{k=1}^{n_j^+} \frac{N \pi_j}{Y_{jk} - \mathbf{b}_k^T \mathbf{x}_j} \right)^2$$

$$\approx \frac{1}{n_j^+} \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{n_j^+} (Y_{jk} - \bar{Y}_{jk} - \bar{Y}_{j+} + \bar{Y}_{++})^2 \equiv \frac{n_j^+}{\hat{d}_k} \quad (35)$$

Notons $\bar{Y}_{jk} = 1/n_j^+ \sum_{k=1}^{n_j^+} Y_{jk}$. Sous l'estimateur groupé de la variance (34), il s'ensuit que

$$\hat{d}_k = \sum_{j=1}^J \frac{n_j^+}{n_j^+} \frac{n_j^+}{n_j^+} - K \sum_{k=1}^K \sum_{j=1}^{n_j^+} \frac{n_j^+}{n_j^+}$$

$$\left(\frac{Y_{jk} - \mathbf{b}_k^T \mathbf{x}_j}{N \pi_j} - \frac{1}{n_j^+} \sum_{k=1}^{n_j^+} \frac{N \pi_j}{Y_{jk} - \mathbf{b}_k^T \mathbf{x}_j} \right)^2$$

$$\approx \frac{1}{n_j^+} \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{n_j^+} (Y_{jk} - \bar{Y}_{jk})^2 \equiv \frac{n_j^+}{\hat{d}_k} \quad (36)$$

En introduisant par substitution ces estimateurs groupés de la variance dans la statistique de Wald (32), nous obtenons

$$W = \frac{1}{\hat{d}_p} \left(\sum_{k=1}^K \frac{1}{n_j^+} \frac{1}{n_j^+} \right)^2 - n_j^+ (\bar{Y}_{++})^2 \quad (37)$$

2.7 Avantages des plans en blocs randomisés

Le principal avantage des PBR est l'élimination de la variation entre les blocs dans l'analyse des effets de traitement. Les unités d'échantillonnage provenant d'une même strate, UPB ou grappe sont plus homogènes que celles provenant de strates, UPB ou grappes différentes, ce qui donne à penser qu'on devrait utiliser des structures d'échantillonnage telles que les strates, les UPB ou les grappes comme variables de bloc dans un PBR (Fienberg et Tanur 1987, 1988). Cette approche assure que chaque strate, UPB ou grappe soit suffisamment représentée dans chaque sous-échantillon. Les intervieweurs peuvent aussi être utilisés comme variables de bloc, puisque la variation des observations due aux effets fixes ou aléatoires d'intervieweurs spécifiés dans le modèle de mesure (1) est alors éliminée. Dans les enquêtes où les intervieweurs recueillent les données par IPAO dans des régions géographiques distinctes, la constitution de blocs d'après les intervieweurs élimine aussi cette variation régionale de la variable cible. La puissance d'une expérience est maximisée si l'on répartit les unités d'échantillonnage proportionnellement entre les traitements sur les blocs, c'est-à-dire $n_j^+ / n_j^+ = n_j^+ / n_j^+$ pour tout $j = 1, \dots, J$ (voir Van den Brakel 2001, chapitre 6). Le meilleur moyen de préserver cette répartition est d'utiliser les intervieweurs comme variables de bloc, puisque le taux de réponse varie considérablement d'un intervieweur à l'autre. La randomisation non contrainte au moyen d'un PRT n'est pas toujours réalisable en pratique. Par exemple, dans le cas des enquêtes par IPAO, où les intervieweurs recueillent les données dans des régions géographiques autour de leur lieu de résidence, il pourrait être nécessaire de restreindre la randomisation des unités d'échantillonnage aux intervieweurs ou aux régions géographiques qui sont des unions de régions d'intervieweurs adjacents pour éviter d'accroître de façon inacceptable la distance que doivent parcourir les intervieweurs. Cette approche mène naturellement au PBR avec les intervieweurs ou les régions comme variables de bloc.

$$(30) \quad \tilde{Y}_K \equiv \left(\sum_{i=1}^n \frac{\pi_i}{Y_{ik}^*} \right)^{-1} \left(\sum_{i=1}^n \frac{\pi_i}{1} \right) = \frac{Y_K}{\sum_{i=1}^n \pi_i}$$

Si $\sum_{i=1}^k 1/\pi_i^* \equiv \hat{N} = N$, alors l'estimateur par le ratio (30)

2.5 Test de Wald

$$(32) \quad W = \sum_K \frac{p_K^{\frac{1}{2}}}{\sum_{k=1}^K \frac{p_k^{\frac{1}{2}}}{I}} - \left(\sum_K \frac{p_K^{\frac{1}{2}}}{\sum_{k=1}^K \frac{p_k^{\frac{1}{2}}}{I}} \right)^2$$

Statistique Canada, N° 12-001-XPB au catalogue

2.6 Estimateurs groupés de la variance

Dans le cas d'un PBR, les n^{++} unités d'échantillonnage de s sont réparties en JK groupes de taille n_{jk}^{+} . Il faut estimer une variance de population S_{jk}^{2+} distincte pour chacun de ces JK sous-échantillons. Si le nombre d'unités expérimentales n_{jk}^{+} disponibles pour l'estimation de ces variances de population devient trop petit, les estimations risquent de devenir instables. Le cas échéant, on peut obtenir des estimations plus stables en regroupant les estimations des variances de population dans les blocs.

$$\mathcal{S}_{\mathcal{E}_{f_j: \mathbf{t}_j}} = \frac{1}{\sum_{n=1}^K \sum_{\mathbf{t}=1}^{K-1} (n)^{f_j-1}} - \frac{1}{\sum_{n=1}^K \sum_{\mathbf{t}=1}^{K-1} \frac{N \pi_{\mathbf{t}}}{n^{f_j+1} - \mathbf{g}_{\mathbf{t}}^*(\mathbf{x}_{\mathbf{t}})}} - \frac{N \pi_{\mathbf{t}}}{\sum_{n=1}^K \sum_{\mathbf{t}=1}^{K-1} \frac{N \pi_{\mathbf{t}}}{n^{f_j+1} - \mathbf{g}_{\mathbf{t}}^*(\mathbf{x}_{\mathbf{t}})}} \quad (33)$$

ou, alternatively,

où B est un vecteur de dimension H contenant les coefficients de régression et les e_i sont les résidus. Dans l'approche assistée par modèle de Särndal et coll. (1992), les valeurs intrinsèques u_i sont considérées chacune comme une réalisation d'un modèle de superpopulation sous-jacent défini par (16). Dans ce cas, les résidus e_i sont des variables aléatoires indépendantes de variance ω_i^2 . Alors, il est nécessaire de connaître tous les ω_i^2 jusqu'à un facteur d'échelle commun; autrement dit, $\omega_i^2 = v_i \omega^2$ avec v_i connu. D'un point de vue strictement axé sur le plan de sondage, proposé par Bethlehem et Keller (1987), il n'est pas nécessaire d'adopter un modèle de superpopulation. Alors, les résidus sont des valeurs intrinsèques fixes des éléments de la population finie et aucune hypothèse de modélisation ne doit être formulée au sujet des résidus. Ici, nous adoptons l'approche assistée par modèle de Särndal. Cela signifie que les espérances par rapport au modèle de mesure, comme en (7) et (10), sont les espérances conditionnelles sachant la réalisation des valeurs intrinsèques $u_i, i = 1, \dots, N$, dans la population finie conformément au modèle de superpopulation (16).

Les coefficients de régression du modèle linéaire (16) dans la population finie sont définis par

$$\mathbf{b} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^N \mathbf{x}_i y_i \right) \quad (17)$$

Les valeurs intrinsèques u_i ne sont pas observables à cause des erreurs de mesure et des effets de traitement. Par conséquent, nous ne pouvons calculer (17), même si nous décombrons entièrement la population finie. Dans le cas d'un recensement complet sous le k^{e} traitement

$$\mathbf{b}^k = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^N \mathbf{x}_i y_{ik} \right) \omega_i^2, \quad k = 1, 2, \dots, K, \quad (18)$$

représente les coefficients de régression en population finie du modèle linéaire (16). Sachant la réalisation de $u_i, i = 1, \dots, N$, l'espérance des coefficients de régression en population finie \mathbf{b}^k par rapport au modèle de l'erreur de mesure est donnée par

$$E^m \mathbf{b}^k = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^N \mathbf{x}_i (u_i + \beta^k + \psi_i) \right) \omega_i^2 \equiv \mathbf{b}^k, \quad (19)$$

Nous pouvons estimer les coefficients de régression en population finie \mathbf{b}^k et \mathbf{b}^k en utilisant les données d'échantillon provenant du sous-échantillon s_k avec l'estimateur d'Horvitz-Thompson :

$$\hat{\mathbf{b}}^k = \left(\sum_{i \in s_k} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i \in s_k} \mathbf{x}_i y_{ik} \right) \omega_i^2, \quad k = 1, 2, \dots, K. \quad (20)$$

$$\mathbf{p}^k \equiv \begin{cases} \frac{1}{n_k} \mathbf{r}_k & \text{si } i \in s_k, \\ \mathbf{0} & \text{si } i \notin s_k \end{cases} \quad (14)$$

et pour un PBR

$$\mathbf{p}^k \equiv \begin{cases} \frac{1}{n_k} \mathbf{r}_k & \text{si } i \in s_k, \\ \mathbf{0} & \text{si } i \notin s_k \end{cases} \quad (15)$$

où \mathbf{r}_k est le vecteur unitaire de dimension K avec le k^{e} élément égal à un et les autres éléments égaux à zéro, et $\mathbf{0}$ représente un vecteur de dimension K de zéros. Les propriétés des vecteurs \mathbf{p}^k sont données en annexe.

Maintenant, puisque s_k peut être considéré comme un échantillon à deux phases, il est vérifié que $E_e(\mathbf{r}_k | s, m) = \mathbf{r}_k$, où E_e et E_e représentent l'espérance par rapport au plan de sondage et au plan d'expérience, respectivement. Donc, sachant m , nous proposons le vecteur $\hat{\mathbf{V}}_{\text{HT}} = (\hat{V}_{1:\text{HT}}, \dots, \hat{V}_{K:\text{HT}})'$ comme estimateur sans biais de \mathbf{V} . Mais alors, $\hat{\mathbf{V}}_{\text{HT}}$ est sans biais pour $E^m \mathbf{V}$.

2.3.2 Estimateur par la régression généralisée

Dans le cas de l'échantillonnage en population finie, on augmente habituellement la précision de l'estimateur d'Horvitz-Thompson, si l'on dispose de données auxiliaires appropriées, au moyen de l'estimateur par la régression généralisée [consulter, par exemple, Bethlehem et Keller (1987) et Särndal, Swensson et Wretman (1992)]. L'estimateur par la régression généralisée nous permet d'intégrer le schéma de pondération de l'enquête courante dans l'analyse des expériences intégrées. Cela pourrait réduire la variance par rapport au plan de sondage, ainsi que le biais dû à la non-réponse sélective et, par conséquent, accroître la précision de l'expérience. Dans le présent contexte, l'estimateur par la régression fondé sur le plan de sondage de covariance de la méthodologie classique des plans d'expérience. En plus des valeurs de la variable de réponse \mathbf{y}_i , nous associons à chaque unité de la population un vecteur d'information auxiliaire \mathbf{x}_i de dimension H . Nous supposons que la moyenne en population finie de ces variables auxiliaires est connue et nous la représentons par \mathbf{X} . Nous supposons aussi que les variables auxiliaires sont des valeurs intrinsèques, qui peuvent être observées sans erreur de mesure et ne sont, par conséquent, pas affectées par les traitements. Si nous suivons l'approche assistée par modèle de Särndal et coll. (1992), nous supposons que les valeurs intrinsèques u_i du modèle de mesure de l'erreur de mesure de la population sont des réalisations indépendantes du modèle de régression linéaire :

L'objectif de l'expérience est de déterminer s'il existe des différences systématiques entre les K moyennes de population de \underline{Y} dues aux K stratégies d'enquête différentes, ou traitements. Nous pouvons pour cela formuler des hypothèses au sujet de

$$E_m(\underline{Y}) = \mathbf{f} \frac{1}{N} \sum_{i=1}^N n_i + \sum_{l=1}^L \frac{1}{N} \sum_{i=1}^N n_i \psi_l + \boldsymbol{\beta}, \quad (10)$$

où l'espérance est prise sur le modèle de l'erreur de mesure. Nous obtenons ainsi les hypothèses suivantes :

$$H_0 : CE_m \underline{Y} = 0, \quad H_1 : CE_m \underline{Y} \neq 0, \quad (11)$$

où \mathbf{C} représente une matrice de dimensions $(K-1) \times K$ avec $K-1$ contrastes et $\mathbf{0}$ représente un vecteur de dimension $K-1$ de zéros. Puisque $\mathbf{C}\mathbf{f} = \mathbf{0}$, il s'ensuit que $CE_m \underline{Y} = \mathbf{C}\boldsymbol{\beta}$ et les hypothèses (11) concernent les effets de traitement représentés par $\boldsymbol{\beta}$ dans le modèle d'erreur de mesure (1). Les contrastes entre les paramètres de population correspondent commodément à ces effets de traitement.

En ce qui concerne les expériences randomisées considérées dans le présent article, il est vérifié que chaque unité expérimentale affectée à un intervieweur l a une probabilité non nulle d'être affecté à chacun des K traitements. Par conséquent, le biais induit dans les estimations des paramètres par les effets fixes d'intervieweurs est le même sous chacun des K traitements et s'annule dans les $K-1$ contrastes entre les K estimations de paramètre.

Nous vérifierons l'hypothèse (11) en estimant $E_m \underline{Y}$ au lieu de $\boldsymbol{\beta}$, en tenant compte du plan d'échantillonnage, du plan d'expérience et de la méthode de pondération appliquée dans l'enquête courante pour estimer les paramètres de population. Pour vérifier (11), nous disposons d'un échantillon probabiliste tiré d'une population finie. Les unités d'échantillonnage (unités expérimentales) sont randomisées sur K sous-échantillons et assignées à l'un des traitements \underline{Y} . À la section 2.3, nous élargirons un estimateur sans biais par rapport au plan de sondage de $E_m \underline{Y}$, que nous notons $\underline{\hat{Y}}$. Par exemple, $\underline{\hat{Y}}$ pourrait être l'estimateur d'Horvitz-Thompson ou l'estimateur par la régression généralisée. Soit \underline{V} la matrice des covariances de $\underline{\hat{Y}}$. Un plan de sondage de la matrice des covariances des $K-1$ contrastes de $\underline{\hat{Y}}$, noté \mathbf{CVC}' , sera établi à la section 2.4. Maintenant, nous pouvons vérifier l'hypothèse (11) au moyen de la statistique de Wald basée sur le plan de sondage qui suit :

$$W = \underline{\hat{Y}}' \mathbf{C}' (\mathbf{CVC}')^{-1} \mathbf{C} \underline{\hat{Y}}, \quad (12)$$

Pour des considérations d'ordre mathématique, nous préférons la matrice de contrastes $\mathbf{C} = (\mathbf{f}; -\mathbf{I})$, où \mathbf{f} est un

vecteur de dimension $K-1$ et \mathbf{I} est la matrice identité de dimensions $(K-1) \times (K-1)$.

2.3 Estimation des effets de traitement

2.3.1 Estimateur d'Horvitz-Thompson

Considérons un échantillon s tiré selon un plan de sondage généralement complexe, qui peut être décrit par les probabilités d'inclusion de premier et de deuxième ordres π_i et π_{ij} de la i^{e} et des i, j^{e} unités d'échantillonnage, respectivement. Dans le cas d'un PRT, l'échantillon s est subdivisé aléatoirement en K sous-échantillons s_k de taille n_k . Si $n_+ = \sum_{k=1}^K n_k$ représente le nombre d'unités d'échantillonnage dans s , alors la probabilité conditionnelle que la i^{e} unité d'échantillonnage soit sélectionnée dans le sous-échantillon s_k est égale à n_k/n_+ . Dans le cas d'un PBR, les unités d'échantillonnage sont, sachant la réalisation de s , subdivisées de façon déterministe en J blocs s_j . Les variables de bloc possibles sont les structures d'échantillonnage telles que les strates, les grappes, les UPE, les intervieweurs et ainsi de suite. Dans chaque bloc, les unités d'échantillonnage sont randomisées sur les K traitements. Soit n_{jk} le nombre d'unités d'échantillonnage dans le bloc j assigné au traitement k . Alors, $n_{j+} = \sum_{k=1}^K n_{jk}$ est la taille du bloc j , $n_{+k} = \sum_{j=1}^J n_{jk}$ est la taille du sous-échantillon s_k et $n_{++} = \sum_{j=1}^J \sum_{k=1}^K n_{jk} = n_{jk}$ est la taille de l'échantillon s . La probabilité conditionnelle que la i^{e} unité d'échantillonnage soit sélectionnée dans le sous-échantillon s_k , sachant que l'échantillon s est sélectionné et que $i \in s_j$, est égale à n_{jk}/n_{j+} .

Nous pouvons considérer chaque sous-échantillon s_k comme un échantillon à deux phases, où les probabilités d'inclusion de premier ordre de l'échantillon de la première phase sont obtenues d'après le plan de sondage et les probabilités d'inclusion conditionnelles de premier ordre de l'échantillon de la deuxième phase sont obtenues d'après le plan d'expérience. De ce point de vue, les probabilités d'inclusion de premier ordre des éléments de s_k sont égales à $\pi_i^* = (n_k/n_+) \pi_i$ pour les PRT et à $\pi_i^* = (n_{jk}/n_{j+}) \pi_i$ pour les PBR si cette i^{e} unité d'échantillonnage est assignée au j^{e} bloc. Il s'ensuit que l'estimateur d'Horvitz-Thompson de \underline{Y}_k , fondé sur les n_{jk} observations obtenues à partir du sous-échantillon s_k , peut être défini par :

$$\underline{\hat{Y}}_{k,HT} = \frac{1}{N} \sum_{i=1}^{n_{++}} \frac{\pi_i^*}{\pi_i^*} = \frac{1}{N} \sum_{i=1}^{n_{++}} \frac{\pi_i}{\pi_i^*}, \quad (13)$$

où les π_i^* sont les vecteurs de dimension K qui décrivent le mécanisme de randomisation du plan expérimental. Pour un PRT, il s'ensuit que

les estimations de divers paramètres de population sous l'approche présentant la même stratégie d'enquête. L'approche présente dans le présent article se résume comme suit. En nous fondant sur les K sous-échantillons, nous établissons un estimateur basé sur le plan de sondage du paramètre de population observé sous chacun des K traitements, ainsi qu'un estimateur fondé sur le plan de sondage de la matrice des covariances des $K - 1$ contrastes entre ces estimations. Cette méthode d'estimation tient compte de la structure probabiliste du plan d'échantillonage, de la randomisation des unités d'échantillonage sur les traitements conformément au plan d'expérience et de la procédure de pondération appliquée dans l'enquête courante pour l'estimation des paramètres cibles. Nous obtenons ainsi une statistique de Wald basée sur le plan de sondage pour tester les hypothèses au sujet des écarts entre les estimations par sondage.

La contribution principale du présent article est d'offrir un cadre général pour la comparaison de K approches d'enquête dans la situation réaliste d'un vrai processus d'enquête par sondage. La sélection aléatoire des unités d'échantillonage à partir d'une population finie cible selon une méthode d'échantillonnage probabiliste est combinée à la randomisation des unités d'échantillonnage sur les divers traitements conformément à un plan d'expérience. Cette façon de procéder facilite la comparaison des effets de diverses approches d'enquête sur les résultats principaux d'une enquête par sondage, ainsi que la généralisation des résultats observés à des populations plus grandes que l'échantillon inclus dans l'expérience. La méthode d'analyse proposée ici généralise l'analyse des expériences à deux traitements intégrés dans les enquêtes par sondage (Van den Brakel et Renssen (1998) et Van den Brakel et Van Berkel (2002)) en l'étendant aux plans en randomisation totale (FRT) et aux plans en blocs randomisés (FBR) avec $K > 2$ traitements. Un résultat important est que l'estimateur fondé sur le plan de sondage de la matrice des covariances possède une structure assez simple, comme si les unités d'échantillonnage étaient tirées avec remise et probabilités de sélection inégales. Par conséquent, la procédure d'estimation de la variance ne nécessite ni les probabilités d'inclusion conjointes ni les covariances fondées sur le plan de sondage entre les estimations sur sous-échantillons, ce qui produit une méthode d'analyse séduisante et assez simple. Un deuxième avantage est que cette méthode permet de tester les hypothèses au sujet des différences entre les estimations par sondage de l'enquête, ce qui facilite l'interprétation des résultats dans de nombreuses applications.

À la section 2, nous présentons une théorie fondée sur le plan de sondage de l'analyse des expériences intégrées. À la section 3, nous présentons une inférence basée sur le plan de sondage pour comparer les paramètres de domaine, où les K traitements sont considérés comme K domaines. Toutefois, l'objectif d'une expérience intégrée est de comparer les estimations du même paramètre sous diverses stratégies d'enquête, ou traitements, alors que, dans le cas des paramètres de domaine, l'objectif est de comparer les sous-échantillons.

La deuxième option consiste à faire à une inférence basée sur le plan de sondage pour comparer les paramètres de domaine, où les K traitements sont considérés comme K domaines. Toutefois, l'objectif d'une expérience intégrée est de comparer les estimations du même paramètre sous diverses stratégies d'enquête, ou traitements, alors que, dans le cas des paramètres de domaine, l'objectif est de comparer les sous-échantillons.

structures d'échantillonnage, comme les strates, les unités primaires d'échantillonnage (LPB), les grappes ou les intervalleurs sont des variables de bloc éventuelles. En général, on assigne à l'enquête courante un grand sous-échantillon qui est utilisé pour la production des publications officielles et sert simultanément de groupe témoin dans l'expérience. L'objectif des expériences intégrées est d'estimer les paramètres de population finie sous les diverses mises en œuvre de l'enquête et de tester les hypothèses au sujet des écarts entre les diverses estimations ainsi obtenues de ces paramètres.

Au premier abord, on pourrait considérer une approche basée sur un modèle classique pour cette analyse. Cependant, puisque les unités expérimentales sont tirées selon un plan d'échantillonnage complexe sans remise à partir d'une population finie, l'application d'une telle approche risque de produire des estimations des paramètres et des variances biaisées par rapport au plan de sondage. Les résultats de l'analyse ne pourraient alors pas être comparés aux estimations des paramètres et des variances de l'enquête ordinaire, ce qui compliquerait l'interprétation des résultats sous les conditions du plan de l'enquête par sondage. Pour rendre l'analyse plus robuste aux écarts par rapport au modèle hypothétique, il faudrait adopter une approche basée sur le plan de sondage qui tient compte de ce dernier.

Avant de présenter notre approche basée sur le plan de sondage, nous mentionnons deux autres options qui, à première vue, semblent correctes. Toutefois, nous argumentons brièvement du fait que ces deux options produisent généralement des résultats invalides. La première est une analyse par régression linéaire fondée sur le plan de sondage qui tient compte du plan d'échantillonnage pour estimer les effets des K traitements introduits dans le modèle de régression et tester les hypothèses à leur sujet. Cependant, cette approche produit facilement des estimations incorrectes des variances par rapport au plan, puisqu'on ignore la randomisation du plan d'expérience. L'objectif analytique principal des expériences intégrées est de comparer les effets de diverses stratégies d'enquête sur les estimations principales produites d'après l'enquête par sondage courante. Or, une analyse par régression linéaire ne permet pas précisément de réaliser cet objectif, puisque, dans le modèle de régression, les effets de traitement ne sont généralement pas égaux aux écarts entre les estimations sur les sous-échantillons.

Analyse d'expériences intégrées dans des plans de sondage complexes

Jan A. van den Brakel et Robert H. Renssen¹

Résumé

Les instituts nationaux de statistique intègrent parfois des expériences dans les enquêtes par sondage courantes afin d'étudier les effets éventuels de diverses techniques d'enquête sur les estimations des paramètres d'une population finie. En vue de l'enquête, nous élaborons une théorie fondée sur le plan de sondage pour analyser des plans en randomisation totale ou des plans en blocs randomisés intégrés dans des plans de sondage complexes généraux. Pour ces deux types de plans d'expérience, nous établissons une statistique de Wald fondée sur le plan de sondage pour l'estimateur d'Horvitz-Thompson et pour l'estimateur par la régression généralisée. Enfin, nous illustrons la théorie au moyen d'une étude en simulation.

Mots clés : Analyse fondée sur le plan de sondage; modèles de l'erreur de mesure; échantillonnage probabiliste; expériences randomisées; superposition.

1. Introduction

Une part de la recherche réalisée dans le domaine de la méthodologie d'enquête consiste à considérer et à évaluer la qualité et l'efficacité des processus d'enquête par sondage mis en place par les instituts nationaux de statistique. L'intégration d'expériences à grande échelle sur le terrain dans les enquêtes par sondage courantes convient particulièrement bien pour quantifier l'effet de diverses mises en œuvre d'une enquête sur le comportement de réponse ou sur les estimations des paramètres de population finie. Ainsi, Statistique Pays-Bas a étudié les effets de diverses conceptions de questionnaire, diverses stratégies d'approche ou diverses lettres préalables à l'enquête sur les deux types de paramètres. Consulter à cet égard Van den Brakel et Renssen (1998), Van den Brakel (2001), ainsi que Van den Brakel et Van Berkel (2002). Les instituts nationaux de statistique évitent généralement de modifier les enquêtes par sondage aussi longtemps que possible afin de produire des séries chronologiques ininterrompues d'estimations des paramètres de population. Toutefois, il est inmanquable qu'ils doivent rajuster les processus d'enquête de temps en temps. Des expériences intégrées peuvent être utilisées pour déceler et quantifier les ruptures de tendance que peuvent causer dans ces séries chronologiques les changements qu'il faut apporter à une enquête par sondage et pour assurer une transition harmonieuse de l'ancien au nouveau plan de sondage. L'exécution en parallèle de l'ancienne approche aux fins des publications courantes si la nouvelle s'avère être un échec.

Les applications d'expériences intégrées décrites dans la littérature avaient pour but d'estimer le biais ou les diverses composantes de la variance dans les modèles de l'erreur de mesure totale. Mahalanobis (1946) a probablement été le premier à lancer l'idée d'intégrer des expériences dans les enquêtes par sondage courantes, sous forme de sous-échantillonnage superposé, pour tester les différences entre intervieweurs sous échantillonnage aléatoire simple et randomisation non contrainte des unités d'échantillonnage entre les intervieweurs. Fellegi (1964), ainsi que Hartley et Rao (1978) ont généralisé cette approche pour estimer les variances de réponse sous des plans de sondage plus complexes et la randomisation contrainte des unités d'échantillonnage. Fienberg et Tanur (1987, 1988, 1989) discutent des dissémbiances et des parallèles entre la théorie des plans expérimentaux et celle de l'échantillonnage en population finie, et de la façon utile et naturelle dont les méthodes statistiques appliquées dans les deux domaines peuvent être combinées pour concevoir et analyser des expériences intégrées. Leur article de 1988 donne un aperçu exhaustif des applications d'expériences intégrées menées dans la littérature.

La situation type considérée dans le présent article est une expérience sur le terrain conçue pour comparer les effets de K mises en œuvre différentes d'une enquête, c'est-à-dire les traitements, sur les estimations des principaux paramètres de population finie d'une enquête courante. À cette fin, un échantillon probabiliste tiré d'une population finie est subdivisé aléatoirement en K sous-échantillons conformément à un plan d'expérience. Chaque sous-échantillon est assigné à l'un des K traitements. Les plans d'expérience considérés sont les plans en randomisation totale (PRT) et les plans en blocs randomisés (PBR), où les

- Krewski, Dewanjli, Wang, Bartlett, Zielinski et Mallick : L'effet des erreurs de couplage d'enregistrements
- Breslow, N.E., et Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2 : *The Design and Analysis of Cohort Studies*. IARC scientific publication No. 82, international agency for research on cancer, Lyon, France.
- Carpenter, M., et Fair, M.E. (1990). *Canadian Epidemiology Research Conference – 1989: Proceedings of Record Linkage Sessions & Workshop*. Ottawa Select Printing, Ottawa.
- Cox, D.R. (1972). Regression models and life tables (avec discussion). *Journal of Royal Statistical Society*, B, 34, 187-220.
- Fair, M.E. (1989). Studies and References Relating to Uses of the Canadian Mortality Data Base. Report from the occupational and environmental health research unit, Division de la Santé, Statistique Canada, Ottawa.
- Felleit, I., et Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Hill, T. (1988). Generalized Iterative Record Linkage system: GIRLS Strategy (Reliacher 2.7). Report from research and general system, Informatics services and development division, Statistique Canada, Ottawa.
- Howe, G.R., et Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.
- Howe, G.R., et Spasoff, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. University of Toronto Press, Toronto.
- Jordan-Simpson, D.A., Fair, M.E. et Poliquin, C. (1990). Étude des exploitants agricoles canadiens : Méthodologie. *Rapports sur la santé*, 2, 141-155.
- Kalbfleisch, J.D., et Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.
- Labossière, G. (1986). Confidentiality and access to data: The Computerized Record Linkage in Health Research, University of Toronto Press, Toronto.
- Mallick, R., Krewski, D., Dewanjli, A. et Zielinski, J.M. (2002). A simulation study of the effect of record linkage errors in cohort mortality data. *Proceedings of International Conference in Recent Advances in Survey Sampling*. Carleton University, Ottawa, à paraître.
- Netter, J., Maynes, E.S. et Ramamanian, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies*. Administration, and Business. Oxford Medical Publications, Oxford.
- Rooos, L.T., Sooden, R. et Jébamani, L. (2001). Un environnement riche en information : La qualité des données des systèmes La qualité des données d'un organisme statistique : Une perspective méthodologique, Statistique Canada, Ottawa.
- Scheuren, F., et Winkler, W.E. (1993). Analyse de regression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.
- Scheuren, W.E. (1997). Analyse de regression des fichiers de données appariés par ordinateur – Partie II. *Techniques d'enquête*, 23, 171-180.
- Singh, A.C., Feder, M., Dunteman, G. et Yu, F. (2001). Protection de la confidentialité et maintien de la qualité des microdonnées à grande diffusion. *Recueil : Symposium 2001, La qualité des données d'un organisme statistique : Une perspective méthodologique*, Statistique Canada, Ottawa.
- Smith, M.E., et Silins, J. (1981). Generalized iterative record linkage system. *Social Statistics Section, Proceedings of the American Statistical Association*, 128-137.
- Sont, W.N., Zielinski, J.M., Ashmore, J.P., Jiang, H., Krewski, D., Fair, M.E., Band, P. et Letourneau, E. (2001). First analysis of cancer incidence and occupational radiation exposure based on the national dose registry of Canada. *American Journal of Epidemiology*, 153, 309-318.
- Winkler, W.E., et Scheuren, F. (1991). How computer matching error effect regression analysis: Exploratory and confirmatory analysis. Census, Washington, D.C.

La présente étude a été financée en partie par une bourse du Conseil national de recherches en sciences et en génie du Canada octroyée à D. Krewski, qui est titulaire à l'heure actuelle de la chaire CRSNG-CRHS-McLaughlin d'évaluation du risque pour la santé des populations à l'Université d'Ottawa. Des versions préliminaires du présent article ont été présentées à l'Annual Joint Meeting de l'American Statistical Association qui s'est tenue à San Francisco du 8 au 12 août 1993 et à l'Assemblée annuelle de la Société statistique du Canada qui s'est tenue à Montréal du 10 au 16 juillet 1995. La version finale a été présentée à la session dédiée à J.N.K. Rao du Symposium 2001 de Statistique Canada qui a eu lieu à Ottawa le 18 octobre 2001. L'auteur principal (D. Krewski) est particulièrement reconnaissant d'avoir été invité à prendre la parole à la session en l'honneur de J.N.K. Rao, qui avait été son directeur de thèse de doctorat il y a de nombreuses années. L'étude a été achevée pendant les séjours de A. Dewanji au Centre McLaughlin d'évaluation du risque pour la santé des populations à titre de chercheur invité durant les étés de 2002 et de 2003.

Bibliographie

- Anderson, T.W. (1974). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, Inc.
- Ardal, S., et Ennis, S. (2001). Enquêtes sur les données : Mise en évidence d'erreurs systématiques dans les bases de données administratives. *Recueil : Symposium 2001. La qualité des données d'un organisme statistique : Une perspective méthodologique*, Statistique Canada, Ottawa.
- Ashmore, J.-P., et Grogan, D. (1985). The national dose registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.
- Ashmore, J.-P., et Davies, B.D. (1989). The national dose registry: A centralized record keeping system for radiation workers in Canada. Dans *Applications of Computer Technology to Radiation Protection*, IAEA-SR-136/58, J. Stephan Institute, Ljubljana, 505-520.
- Ashmore, J.-P., Krewski, D. and Zielinski, J.M. (1997). Protocol for a cohort mortality study of occupational radiation exposure based on the national dose registry of Canada. *European Journal of Cancer*, 33, S10-S21.
- Ashmore, J.-P., Krewski, D., Zielinski, J.M., Jiang, H., Semenciw, R. et Lévesque, E. (1998). First analysis of occupational radiation mortality based on the national dose registry of Canada. *American Journal of Epidemiology*, 148, S64-S74.
- Bartlett, S., Krewski, D., Wang, Y. et Zielinski, J.M. (1993). Evaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisés. *Techniques d'enquête*, 19, 3-13.
- Belin, T.R., et Rubin, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.
- Breslow, N.E., Lubin, J.H. et Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78, 1-12.
- Statistique Canada, N° 12-001-XPB au catalogue
- par α la probabilité d'une erreur de couplage (de l'un ou l'autre type), nous pouvons écrire les taux de résultats faussement positifs et de résultats faussement négatifs, p_f^+ et p_f^- , sous la forme $\alpha P[T \leq \tau_j]$ et $\alpha P[T > \tau_j]$, respectivement. En particulier, $p_f^+ = \alpha P[\tau_{j-1} < T \leq \tau_j]$, où τ_{j-1} est la limite inférieure d'âge pour le j^{e} état, et $p_f^- = p_{j-1}^+$. Par conséquent, les taux de résultats faussement positifs peuvent être supérieurs aux taux de résultats faussement négatifs pour les groupes d'âge avancé, l'inverse se produisant pour les groupes d'âge plus jeune. Si l'on suppose que le profil de taille est le même pour les D_j et A_j , certains termes s'annulent dans le calcul de $E[\Delta e_j]$ dans (19) et dans celui de $E[\Delta d_j]$ dans (15). Cet effet d'annulation réduira les biais attendus dans le RSM et dans les paramètres de régression du risque donnés par (23) et (38), respectivement.
- Bien que nous ayons considéré uniquement la mortalité toutes causes combinées dans le présent article, la mortalité par cause peut être étudiée en apportant des modifications simples aux définitions de D_j^+ , D_j^- et D_j^0 . Ces ensembles devraient alors ne tenir compte que des décès dus à la cause particulière étudiée. Par conséquent, d_j^+ et e_j^+ devraient représenter, respectivement, les nombres observé et attendu de décès du type spécifié dans S_j . Dans (1) et (2), la fonction de risque devrait avoir trait au type spécifique de décès, avec $\lambda^*(u)$ le taux de risque par cause de base correspondant. Enfin, à la section 2, l'indicateur δ_j devrait indiquer le type spécifique de décès.
- Les résultats analytiques qui précèdent fournissent d'importants éclaircissements sur les effets des erreurs de couplage dans les études-cohorte de la mortalité, mais il est important d'examiner ce genre d'effets dans des conditions aussi proches que possible de celles rencontrées en pratique. À cette fin, nous avons réalisé une étude en simulation informatisée fondée sur des données réelles provenant du Fichier d'osmétrique national du Canada, dans laquelle nous avons introduit des couplages incorrects et des non-couplages incorrects avec probabilités connues pour évaluer plus en profondeur l'effet des erreurs de couplage sur les estimations du risque de cancer (Maillick, Krewski, Dewanji et Zielinski 2002). Les résultats de cette simulation correspondent aux méthodes classiques d'analyse des données sur la mortalité des cohortes. Des travaux de recherche dans ce domaine sont en cours.

$$\Delta\beta \approx \frac{\sum_j \Delta d_{jj} + e^\beta \sum_j \Delta e_j}{e^\beta \sum_j e_j} \quad (43)$$

Puisque le RSM = $e^\beta = \sum_j d_{jj} / \sum_j e_j$, avec $\Delta\beta = \Delta\text{RSM}/\text{RSM}$ ici, nous obtenons

$$\Delta\beta \approx \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j} \quad (44)$$

Donc, l'expression (44) peut être considérée comme un cas particulier de (22).

Les résultats qui précèdent indiquent que les résultats

faussement positifs ainsi que faussement négatifs introduisent un biais et une variation supplémentaire dans les

estimations des paramètres de régression du risque relatif.

La seule contribution négative à cette variance supplémentaire est le premier terme de (32) (voir $\Theta_{jj'}$). En

utilisant le même argument qu'à la section 4.1, il s'ensuit

que cette variance supplémentaire est strictement positive.

5. Conclusion

Le couplage d'enregistrements est maintenant une

technique bien établie dans le contexte des études épidémiologiques des risques pour la santé des populations. En cou-

plant l'information sur les expositions des individus provenant d'une base de données, il est possible de construire de grandes bases de

données informatives sur les risques que courent les populations et les sous-groupes de population. Le succès de

ce genre d'études dépend, en grande partie, de la qualité des deux bases de données que l'on couple, y compris la

quantité d'information sur les identificateurs personnels utilisés pour coupler les individus représentés dans les deux

bases de données. Dans la plupart des études, l'exactitude du couplage est examinée en estimant les taux de faux

couplages (résultats faussement positifs) et de faux non-couplages (résultats faussement négatifs) associés au

processus de couplage. En pratique, on procède habituellement au tirage d'un échantillon d'enregistrements couplés et

non couplés, puis on détermine l'exactitude des couplages dans l'échantillon en se servant de données auxiliaires provenant d'autres sources.

Bien que le CIE soit utilisé depuis un certain temps dans les études-cohorte de mortalité, l'effet des erreurs de cou-

plages sur la fiabilité des inférences statistiques faites d'après ce genre d'études n'a pas fait l'objet d'un examen détaillé.

Les résultats théoriques présentés dans le présent article visent à combler cette lacune. Ces résultats montrent qu'en

plus d'accroître le nombre observé de décès, les résultats

de décès. Inversement, les résultats faussement négatifs

accroissent le nombre attendu de décès et réduisent le

faussement positifs ont tendance à réduire le nombre attendu

de décès. Bien que nous émettions l'hypothèse simplifiante que

$t_{ij}^0 = t_{ij}^1$, il est possible d'établir les expressions pertinentes du biais et de la variabilité supplémentaire sans le faire;

cependant, les expressions sont trop complexes pour fournir

des éclaircissements supplémentaires sur les effets des erreurs de couplage. Il en est également ainsi de l'hypothèse

selon laquelle $p_{jj'}^N = p_{jj'}^T$. La définition de A_j pour le ou les

états correspondant à la dernière tranche d'âge, qui est

habituellement ouverte jusqu'à l'infini du côté droit, pose un

problème technique. Dans ces états, l'hypothèse que $t_{ij}^0 = t_{ij}^1$

est problématique si la probabilité de mourir dans cette

dernière tranche d'âge est appréciable. On peut contourner

le problème en supposant que la durée de vie humaine a une

limite supérieure finie.

Comme nous en discutons à la section 3.1, les résultats

faussement positifs surviennent principalement lorsqu'un

individu en vie à la fin de la période de suivi est couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

sonne décédée dans l'un des états S_j peut être couple in-

correctement à une personne décédée. Cependant, une per-

En utilisant les équations (25) à (32), nous pouvons approximer la variance de la différence relative $\Delta RSM/RSM$ au moyen du deuxième membre de (24). Nous pouvons tirer deux conclusions des équations (23) et (24). En premier lieu, les erreurs de couplage peuvent introduire un biais dans l'estimation du RSM. En deuxième lieu, les deux types d'erreurs de couplage introduisent une variation supplémentaire dans les estimations du RSM. Notons que le premier terme de (32) est dominé par le premier terme de (29) pour

$p_j^* < 0,5$, et que le terme de covariance négatif (28) est dominé dans le calcul de la variance dans (25). Par conséquent, la variance supplémentaire (24) est strictement positive, puisque les taux de résultats faussement positifs et de résultats faussement négatifs sont tous deux positifs.

4.2 Paramètres de régression du risque relatif

Pour déterminer l'effet des erreurs de couplage sur les estimations des paramètres de régression, considérons d'abord le modèle général de régression du risque relatif (2). En remplaçant dans la fonction de log-vraisemblance (7) les nombres observés et attendu de décès d_j^H et e_j par les nombres observés et attendu de décès en présence d'erreurs de couplage d_j^H et e_j^f , nous obtenons

$$\log L = \sum_{j=1}^J \{d_j^H \log(\gamma\{\beta^f z_j\}) - \gamma\{\beta^f z_j\} e_j^f\}. \quad (33)$$

Soit β et β^f les estimations du maximum de vraisemblance de β fondées sur $\{d_j^H, e_j\}$ et $\{d_j^H, e_j^f\}$, respectivement. L'équation de score (9) peut s'écrire sous la forme

$$\sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} = 0. \quad (34)$$

En supposant que $\Delta\beta = \beta - \beta^f$ est faible, un développement en série de premier ordre de $\exp\{V(\beta)\}$ autour de β donne

$$\exp\{V(\beta^f)\} \approx \exp\{V(\beta)\} + \exp\{V(\beta)\} \frac{\partial}{\partial \beta} \Delta\beta. \quad (35)$$

où $V(\beta) = V(\beta)$ et $\partial V(\beta)/\partial \beta$ est $\partial V(\beta)/\partial \beta$ évalué à $\beta = \beta$. En introduisant (35) par substitution dans (34), nous obtenons

$$\sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} - \exp\{V(\beta)\} \frac{\partial}{\partial \beta} \Delta\beta + \gamma\{\beta^f z_j\} \Delta e_j - \gamma\{\beta^f z_j\} e_j^f \frac{\partial}{\partial \beta} \Delta\beta \approx 0. \quad (36)$$

En utilisant (9), la première somme dans (36) est nulle. Par conséquent, puisque $\Delta e_j \Delta \beta$ est faible, $\Delta \beta$ peut être approximé par

$$\Delta \beta = \left(\sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} e_j^f \frac{\partial}{\partial \beta} \right)^{-1} \sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} \Delta e_j. \quad (37)$$

Il découle de (37) que

$$E[\Delta \beta] \approx \left(\sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} e_j^f \frac{\partial}{\partial \beta} \right)^{-1} \sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} \alpha_j, \quad (38)$$

où $\alpha_j = E[\Delta d_j^H] + \gamma\{\beta^f z_j\} E[\Delta e_j]$, qui peut être calculé d'après (15) et (19). En outre,

$$V[\Delta \beta] \approx \left(\sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} e_j^f \frac{\partial}{\partial \beta} \right)^{-1} \sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} \left(\sum_{j=1}^J \frac{\partial}{\partial \beta} \log \{V(\beta)\} \gamma\{\beta^f z_j\} e_j^f \frac{\partial}{\partial \beta} \right)^{-1} \quad (39)$$

avec

où $X' = (z_1^f, \dots, z_J^f)$, $\Delta D' = (\Delta d_{11}, \dots, \Delta d_{JJ})$, $W = \text{diag}(\exp(z_1^f \beta), \dots, \exp(z_J^f \beta))$, et $\Delta W' = (\exp(z_1^f \beta) \Delta e_1, \dots, \exp(z_J^f \beta) \Delta e_J)$. Notons que la matrice de poids W est la matrice d'information de Fisher pour β . Il découle de (38) que

$$E[\Delta \beta] \approx (X' W X)^{-1} X' W \Pi, \quad (41)$$

où $\Pi' = (\pi_1, \dots, \pi_J)$ avec π_j identique à α_j , mais $\gamma\{\beta^f z_j\}$ remplacé par $\exp(z_j^f \beta)$. En outre,

$$V[\Delta \beta] \approx (X' W X)^{-1} X' \Psi X (X' W X)^{-1}, \quad (42)$$

où Ψ est la matrice des Θ_j^f avec $\gamma\{\beta^f z_j\}$ remplacé par $\exp(z_j^f \beta)$. Notons que les expressions (40) à (42) sont des cas particuliers des expressions (37) à (39), respectivement, avec une seule covariable $z_j = 1$, $X' W X = e^{\beta} \sum_j e_j^f$, $X' \Delta D = \sum_j d_j^H$ et $X' \Delta W = e^{\beta} \sum_j \Delta e_j$. Dans ce cas,

général, la diminution de p_j^f donnera lieu à une augmentation de p_N^f , et inversement (voir la section 5 pour une discussion plus approfondie de ce point). Bien que ces taux d'erreurs soient indépendants du modèle de régression du risque relatif sous-jacent γ donné par (2), l'erreur quadratique moyenne obtenue par combinaison des termes d'espérance et de variance ne peut être minimisée sans qu'on spécifie le risque de base $\mathcal{N}^*(u)$, qui figure dans T_λ .

4. L'effet des erreurs de couplage sur les estimations des RSM et des coefficients de régression

4.1 Ratios standardisés de mortalité

Pour déterminer l'effet des erreurs de couplage sur les RSM, nous remplaçons les nombres observé et attendu réels de décès d_{jj}^f et e_j par les nombres observé et attendu dans l'expression $\text{RSM} = \sum d_{jj}^f / \sum e_j$. En représentant par RSM_L les ratios standardisés de mortalité en présence d'erreurs de couplage, nous obtenons

$$\text{RSM}_L = \text{RSM} \left[1 + \frac{\sum \Delta d_{jj}^f}{\sum \Delta e_j} \right] / \left[1 - \frac{\sum \Delta e_j}{\sum e_j} \right]. \quad (21)$$

Il découle des équations (10) à (14) que les résultats faussément positifs feront augmenter le RSM, tandis que les résultats faussement négatifs le feront diminuer. En utilisant un développement en série de premier ordre de Taylor comme approximation de RSM_L autour de RSM , la différence $\Delta \text{RSM} = \text{RSM}_L - \text{RSM}$ peut s'exprimer sous la forme

$$\frac{\Delta \text{RSM}}{\text{RSM}} = \frac{\sum_j \Delta d_{jj}^f}{\sum_j d_{jj}^f} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \quad (22)$$

Alors, la moyenne et la variance de la différence relative de ΔRSM peuvent être approximées par

$$E \left[\frac{\Delta \text{RSM}}{\text{RSM}} \right] \approx \frac{\sum_j E[\Delta d_{jj}^f]}{\sum_j E[d_{jj}^f]} + \frac{\sum_j E[e_j]}{\sum_j e_j} \quad (23)$$

et

$$V \left[\frac{\Delta \text{RSM}}{\text{RSM}} \right] \approx \left(\sum_j d_{jj}^f \right)^{-2} V \left[\sum_j \Delta d_{jj}^f \right] + \left(\sum_j e_j \right)^{-2} V \left[\sum_j \Delta e_j \right] + 2 \left(\sum_j d_{jj}^f \right)^{-1} \left(\sum_j e_j \right)^{-1} \text{Cov} \left[\sum_j \Delta d_{jj}^f, \sum_j \Delta e_j \right], \quad (24)$$

respectivement. Il est facile de calculer le deuxième membre de (23) en utilisant (15) et (19). Pour calculer le deuxième membre de (24), notons que

$$\begin{aligned} \text{Cov}[\Delta e_j, \Delta e_{j'}] &= - \sum_{i \in A_j \cap A_{j'}} d_p^f (1 - p_p^f) T_\lambda(i, j) T_\lambda(i, j') \\ &\quad + \sum_{i \in A_j \cap D_{j'}} d_p^f d_{j'}^f T_\lambda(i, j) T_\lambda(i, j') \\ &\quad + \sum_{i \in D_j \cap D_{j'}} d_N^f (1 - p_N^f) T_\lambda(i, j) T_\lambda(i, j'), \end{aligned} \quad (28)$$

Sans perte de généralité, supposons, pour $j < j'$, que $t_{ij}^f \leq t_{ij'}^f$ pour le même individu i (en vie ou décédé) dans S_j^f et $S_{j'}^f$; autrement dit, le moment de l'entrée dans S_j^f est identique ou antérieur à celui de l'entrée dans $S_{j'}^f$. Nous

$$\begin{aligned} \text{Cov}[\Delta d_{jj}^f, \Delta d_{jj'}^f] &= \sum_{j' \neq j} \text{Cov}[\Delta d_{jj}^f, \Delta e_{j'}] + \sum_{j' < j} \text{Cov}[\Delta d_{jj}^f, \Delta e_{j'}]. \end{aligned} \quad (27)$$

$$V \left[\sum_j \Delta e_j \right] = \sum_j V[\Delta e_j] + 2 \sum_{j' < j} \text{Cov}[\Delta e_j, \Delta e_{j'}], \quad (26)$$

$$V \left[\sum_j \Delta d_{jj}^f \right] = \sum_j V[\Delta d_{jj}^f] + 2 \sum_{j' < j} \text{Cov}[\Delta d_{jj}^f, \Delta d_{jj'}^f], \quad (25)$$

$$\begin{aligned} \text{Cov}[\Delta d_{jj}^f, \Delta e_j] &= \sum_{i \in A_j} d_p^f (1 - p_p^f) T_\lambda(i, j) \\ &\quad + \sum_{i \in A_j \cap D_{j'}} d_p^f d_{j'}^f T_\lambda(i, j) T_\lambda(i, j') \\ &\quad + \sum_{i \in D_j \cap D_{j'}} d_N^f (1 - p_N^f) T_\lambda(i, j) T_\lambda(i, j'), \end{aligned} \quad (30)$$

$$\begin{aligned} \text{Cov}[\Delta d_{jj}^f, \Delta e_{j'}] &= \sum_{i \in A_{j'}} d_p^f (1 - p_p^f) T_\lambda(i, j') \\ &\quad + \sum_{i \in A_j \cap D_{j'}} d_p^f d_{j'}^f T_\lambda(i, j) T_\lambda(i, j') \\ &\quad + \sum_{i \in D_j \cap D_{j'}} d_N^f (1 - p_N^f) T_\lambda(i, j) T_\lambda(i, j'), \end{aligned} \quad (31)$$

$$\begin{aligned} \text{Cov}[\Delta d_{jj'}^f, \Delta e_j] &= - \sum_{i \in A_j \cap A_{j'}} d_p^f d_{j'}^f T_\lambda(i, j) T_\lambda(i, j') \\ &\quad + \sum_{i \in A_j \cap D_{j'}} d_p^f d_{j'}^f T_\lambda(i, j) T_\lambda(i, j') \\ &\quad + \sum_{i \in D_j \cap D_{j'}} d_N^f (1 - p_N^f) T_\lambda(i, j) T_\lambda(i, j'), \end{aligned} \quad (32)$$

la cohorte d'intérêt. Lorsque les enregistrements sur l'exposition d'une personne en vie sont associés incorrectement à ceux d'une personne décédée, cette dernière n'appartient habituellement pas à la cohorte. Donc, la contribution de personnes-années à risque de la personne qui demeure en vie cessera prématurément dans l'année du décès présumé; les personnes-années à risque perdues correspondent à la période écoulée de l'année du décès présumé jusqu'à la fin du suivi. Par ailleurs, si les enregistrements sur l'exposition d'un individu décédé sont associés incorrectement à ceux d'une personne en vie, la contribution de personnes-années à risque de cet individu inclura une période supplémentaire s'étendant de l'année réelle du décès jusqu'à la fin du suivi. Par conséquent, les résultats faussement positifs réduiront le nombre de personnes-années à risque dans la cohorte et les résultats faussement négatifs l'augmenteront.

3.2 Espérances et variances des différences dans les nombres observé et attendu de décès

L'effet des erreurs de couplage sur les nombres observé et attendu de décès dépend des taux de résultats faussement

positifs et faussement négatifs. Soit p_j^f et p_j^n les taux de résultats faussement positifs et de résultats faussement négatifs, respectivement, dans l'état S_j , pour $j = 1, \dots, J$, que l'on suppose être constants dans S_j et les mêmes pour tous les individus dans A_j et D_j , respectivement. Cette hypothèse est raisonnable si les individus qui se trouvent dans le même état sont très homogènes, particulièrement en ce qui concerne des attributs tels que la qualité des identificateurs personnels, qui influent sur les taux d'erreurs de couplage. Bien que cette hypothèse idéaliste soit peu susceptible d'être entièrement satisfaisante en pratique, elle simplifie considérablement l'évaluation subséquente des effets des erreurs de couplage. Formellement, $p_j^f(p_j^n)$ est la probabilité conditionnelle qu'un individu compris dans $A_j(D_j)$ soit étiqueté comme étant décédé (en vie) dans l'état S_j . Autrement dit, $p_j^f = P\{i \in D_j^f | i \in A_j\}$ et $p_j^n = P\{i \in A_j^n | i \in D_j\}$.

Soit a_j, d_j, a_j^n et d_j^f le nombre d'individus dans A_j, D_j, A_j^n et D_j^f , respectivement. Alors, notons que d_j^f suit une loi binomiale(a_j, p_j^f) et que a_j^n suit une loi binomiale(d_j, p_j^n). En outre, d_j^f suit une loi binomiale(a_j, p_j^f), ou d_j^f est la probabilité conditionnelle qu'un individu compris dans A_j soit étiqueté comme étant décédé dans l'état S_j . Autrement dit, $p_j^f = P\{i \in D_j^f | i \in A_j\}$. De la même façon, a_j^n suit une loi binomiale(d_j^n, p_j^n), où p_j^n est la probabilité conditionnelle qu'un individu compris dans A_j soit étiqueté comme étant décédé dans l'état S_j . Autrement dit, $p_j^n = P\{i \in A_j^n | i \in D_j\}$. Bien qu'il n'existe pas de relation sans importance entre p_j^n et p_j^f en

général, il est raisonnable de supposer que $p_j^n = p_j^f$ dans le contexte des erreurs de couplage.

En supposant que les erreurs de couplage associées à divers individus sont indépendantes, l'espérance et la variance de la différence dans le nombre observé de décès dans l'état S_j , donnée par Δd_j dans (14), sont

$$E[\Delta d_j] = E[d_j^f] - E[a_j^n] = a_j p_j^f - d_j p_j^n \quad (15)$$

et

$$V[\Delta d_j] = V[d_j^f] + V[a_j^n] \quad (16)$$

$$= a_j p_j^f (1 - p_j^f) + d_j p_j^n (1 - p_j^n).$$

Puisque A_j et D_j sont constitués d'ensembles différents d'individus, d_j^f et a_j^n sont indépendants.

De la même façon, l'espérance et la variance de la différence dans le nombre attendu de décès dans l'état S_j , donnée par Δe_j dans (11), peuvent être calculées comme suit. À cette fin, il est commode d'écrire e_j^f et e_j^n en fonction des variables indicatrices qui suivent. Pour $i \in A_j$, définissons $\xi_{ij} = I\{i \in D_j^f\}$ et $\xi_{ij}^n = I\{i \in D_j^n\}$. En outre, pour $i \in D_j$, définissons $\psi_{ij} = I\{i \in A_j^n\}$. Alors, il découle de (12) et des définitions de D_j^f et A_j^n que

$$e_j^f = \sum_{i \in A_j} \xi_{ij} T_{ij}(i, j) \quad (17)$$

et

$$e_j^n = \sum_{i \in D_j} \psi_{ij} T_{ij}(i, j). \quad (18)$$

En particulier, nous pouvons écrire $d_j^f = \sum_{i \in A_j} \xi_{ij}^f$ et $a_j^n = \sum_{i \in D_j} \psi_{ij}^n$, qui sont utiles pour établir (15) et (16). D'après (17) et (18), nous obtenons

$$E[\Delta e_j] = E[e_j^f] - E[e_j^n] \\ = p_j^f \sum_{i \in A_j} T_{ij}(i, j) - \sum_{i \in D_j} T_{ij}(i, j), \quad (19)$$

et

$$V[\Delta e_j] = V[e_j^f] + V[e_j^n] \\ = p_j^f (1 - p_j^f) \sum_{i \in A_j} T_{ij}^2(i, j) \\ + p_j^n (1 - p_j^n) \sum_{i \in D_j} T_{ij}^2(i, j), \quad (20)$$

Puisque A_j et D_j sont constitués d'ensembles différents d'individus,

Les résultats (15)–(16) et (19)–(20) indiquent que les erreurs de couplage d'enregistrements introduisent un biais attendu de décès. Minimiser les termes de variance dans (16) et (20) est difficile, puisque les deux taux d'erreurs p_j^f et p_j^n ne sont pas fonctionnellement indépendants. En

incorrectement désigné comme étant décédé et un résultat faussement négatif survient quand un membre décédé de la cohorte est considéré comme étant en vie. Plus précisément, pour le développement mathématique qui suit, un résultat faussement positif survient dans un état particulier quand un individu qui demeure en vie pendant tout le temps où il se trouve dans cet état est incorrectement étiqueté comme étant décédé dans cet état. Parallelement, un résultat faussement négatif survient dans un état particulier quand un membre de la cohorte qui est décédé avant d'atteindre cet état ou pendant qu'il se trouvait dans cet état est considéré comme étant en vie en étant dans cet état. Dans un état donné, les résultats faussement positifs et faussement négatifs représentent donc des cas particuliers de l'erreur de classification discutée par Anderson (1974, chapitre 6.2.1). À la présente section, nous examinons l'effet de ces deux types d'erreurs de couplage sur les nombres observés et attendu de décès, respectivement. À cet fin, nous commençons par définir des jeux d'indices dans les divers états que nous utiliserons pour représenter les ensembles d'enregistrements correctement

3.1 Erreurs de couplage

Soit A_j et D_j l'ensemble d'étiquettes pour les membres de la cohorte qui demeurent en vie dans l'état S_j , et pour ceux qui sont décédés dans l'état S_j , respectivement. Soit D_{jj} le sous-ensemble de D_j correspondant aux personnes qui sont décédées dans l'état S_j . Soit A_j^f , D_j^f et D_{jj}^f les ensembles correspondants à la présence d'erreurs de couplage. Définissons en outre D_j^f comme étant l'ensemble d'étiquettes des individus en vie dans l'état S_j (c'est-à-dire dans A_j), mais étiquetés comme étant décédés dans l'état S_j , c'est-à-dire correspondant aux résultats faussement positifs dans S_j . De la même façon, A_j^f est l'ensemble d'individus décédés dans l'état S_j (c'est-à-dire dans D_j), mais étiquetés comme en étant en vie dans l'état S_j , c'est-à-dire correspondant aux résultats faussement positifs dans S_j . Représentons aussi par D_{jj}^f le sous-ensemble de D_{jj}^f correspondant aux individus étiquetés comme étant décédés dans l'état S_j et, pareillement, par A_j^f le sous-ensemble d'individus de A_j^f qui sont décédés dans l'état S_j (c'est-à-dire dans D_{jj}^f). Ces ensembles satisfont aux relations $A_j^f = (A_j - D_j^f) \cup A_j^f$, $D_j^f = (D_j - A_j^f) \cup D_j^f$ et $D_{jj}^f = (D_{jj} - A_{jj}^f) \cup D_{jj}^f$. L'effet des erreurs de couplage sur la fonction de vraisemblance donnée par (7) peut être décrit comme suit. Soit t_j^f le temps auquel le i^e individu entre, réellement ou par erreur de couplage, dans le j^e état S_j . De même, t_j^f représente le moment du décès (s'il a lieu, réellement ou par erreur de la sortie de l'état S_j , réellement ou par erreur du couplage. Notons que, si t_j^f existe, il est inférieur ou égal

à t_j^f . Par souci de simplicité, supposons que t_j^f , s'il existe, est égal à t_j^f ; autrement dit, tous les décès qui surviennent dans un état particulier le font au moment correspondant de l'entrée dans cet état. Bien que cette hypothèse produise une sous-estimation du nombre attendu de décès, aux fins de l'étude du biais, elle n'est peut-être pas si contestable. Le fait de supposer que tous les décès surviennent au moment de la sortie des états correspondants offre aussi une simplification comparable. Partant de (8) et de la décomposition de A_j^f , nous pouvons écrire le nombre attendu de décès e_j^f dans S_j en présence d'erreurs de couplage sous la forme

$$e_j^f = \sum_{i \in A_j^f} \int_{t_0^f}^{t_j^f} \chi^*(n) du + \sum_{i \in A_j^f} \int_{t_0^f}^{t_j^f} \chi^*(n) du - \sum_{i \in D_{jj}^f} \int_{t_0^f}^{t_j^f} \chi^*(n) du - \Delta e_j, \quad (10)$$

$$e_j = \sum_{i \in A_j} \int_{t_0^f}^{t_j^f} \chi^*(n) du, \text{ et } \Delta e_j = e_j^f - e_j^f \quad (11)$$

$$e_j^f = \sum_{i \in A_j^f} \int_{t_0^f}^{t_j^f} \chi^*(n) du \text{ et } e_j^f = \sum_{i \in A_j^f} \int_{t_0^f}^{t_j^f} \chi^*(n) du. \quad (12)$$

Pour simplifier la notation, écrivons $T_\lambda(i, j)$ pour $\int_{t_0^f}^{t_j^f} \chi^*(n) du$ dans la suite. Le terme Δe_j représente le biais introduit par les erreurs de couplage dans le nombre attendu de décès dans le j^e état. Il découle de (11) que les résultats faussement positifs ont tendance à réduire le nombre attendu de décès et que les résultats faussement négatifs ont tendance à l'augmenter. En utilisant la décomposition de D_{jj}^f , nous pouvons écrire le nombre observé de décès d_{jj}^f en présence d'erreurs de couplage comme suit

$$d_{jj}^f = d_{jj} + \Delta d_{jj}, \quad (13)$$

$$\Delta d_{jj} = d_{jj}^f - a_{jj}^f, \quad (14)$$

avec d_{jj} , d_{jj}^f et a_{jj}^f le nombre d'individus dans les ensembles D_{jj} , D_{jj}^f et A_{jj}^f , respectivement. Le terme Δd_{jj} représente la variation du nombre observé de décès dans le j^e état due aux erreurs de couplage. Il découle de (13) et de (14) que les résultats faussement positifs font augmenter le nombre observé de décès et que les résultats faussement négatifs le réduisent. Le statut vital est souvent déterminé par couplage des données sur la cohorte étudiée à celles de la BCDM, dont l'effectif est généralement beaucoup plus grand que celui de

de ces paramètres. À la section 5, nous présentons nos conclusions.

2. Modèles de régression du risque relatif

Les méthodes statistiques d'analyse des données provenant d'études-cohorte de la mortalité sont bien établies (Breslow et Day 1987). L'objectif principal de ce genre d'analyse est de déterminer si l'exposition à l'agent d'intérêt augmente le taux de mortalité chez les membres de la cohorte. La mortalité est caractérisée par la fonction de risque, qui précise le taux de mortalité sous forme de fonction du temps. Si nous représentons par T le moment du décès, la fonction de risque au temps u se définit formellement comme suit

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{\Pr\{u \leq T < u + \Delta u | T \geq u\}}{\Delta u}. \quad (1)$$

Soit $\lambda_i(u)$ la fonction de risque pour une cause particulière de décès au temps u pour l'individu $i = 1, \dots, N$ dans une cohorte de taille N , et soit $\mathbf{z}_i(u)$ un vecteur correspondant de covariables propres à cet individu. Nous supposons que ces covariables ont pour effet de modifier le risque de base $\lambda(u)$ conformément au modèle de régression du risque relatif

$$\lambda_i(u) = \lambda^*(u) \gamma\{\beta' \mathbf{z}_i(u)\}, \quad (2)$$

où γ est une fonction positive des covariables et β est un vecteur de paramètres de régression.

Deux cas particuliers du modèle général de régression du risque relatif présentent un intérêt sont les modèles multiplicatif et additif de régression du risque. Définissons la fonction γ figurant dans (2) par

$$\log \gamma(z) = \frac{p}{(1+z)^p - 1}. \quad (3)$$

Quand $p = 1$, le modèle général de régression du risque relatif se réduit au modèle multiplicatif de régression du risque

$$\lambda_i(u) = \lambda^*(u) \exp\{\beta' \mathbf{z}_i(u)\}, \quad (4)$$

Ce modèle à risques proportionnels, qui a été introduit par Cox (1972), est d'usage très répandu en analyse des données sur la mortalité (Kalbfleisch et Prentice 1980). Le modèle additif de régression du risque

$$\lambda_i(u) = \lambda^*(u) + \beta' \mathbf{z}_i(u) \quad (5)$$

survient en tant que cas limite quand $p \rightarrow 0$. Soit t_i^0 et t_i^1 l'âge au moment de l'entrée dans l'étude et l'âge au moment de la perte de vue (due à l'abandon par le sujet, à l'interruption de l'étude ou au décès) du i^{e} sujet de

la cohorte, respectivement. Soit $\delta_i = 1$ ou 0 , selon que le i^{e} sujet est ou n'est pas décédé au moment de la perte de vue. La fonction de log-vraisemblance fondée sur le modèle du risque relatif (2) peut s'écrire

$$\log L = \sum_{i=1}^N \left\{ \delta_i \log \gamma\{\beta' \mathbf{z}_i(t_i^1)\} - \int_{t_i^0}^{t_i^1} \lambda^*(u) \gamma\{\beta' \mathbf{z}_i(u)\} du \right\}. \quad (6)$$

Lorsqu'il n'existe qu'une covariable $z_i(u) \equiv 1$, l'estimation du maximum de vraisemblance de $\theta = \exp\{\beta\}$ se réduit au ratio standardisé de mortalité $\text{RSM} = \text{OBS}/\text{ATT}$, où $\text{OBS} = \sum_{i=1}^N \delta_i$ et $\text{ATT} = \sum_{i=1}^N e_i$ sont les nombres observés et attendus de décès, respectivement, avec $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u) du$.

La maximisation de la fonction de vraisemblance (6) peut donner lieu à des calculs fastidieux dans le cas d'échantillons de grande taille. Breslow, Lubin et Langholz (1983) simplifient cette fonction en supposant que les covariables prennent des valeurs constantes dans les états par lesquels passe un sujet durant le cours de l'étude. Ces états sont définis par des classifications croisées des covariables d'intérêt. Plus précisément, supposons qu'il existe J états de ce genre $\{S_j; j = 1, \dots, J\}$, tels que $\mathbf{z}_j(u) = \mathbf{z}$ chaque fois que le i^{e} sujet se trouve dans l'état S_j au temps u . Ces états sont mutuellement exclusifs et exhaustifs, si bien que, à tout temps u , chaque membre de la cohorte se trouvera dans un état, et uniquement un. La fonction de log-vraisemblance (6) peut s'écrire

$$\log L = \sum_{j=1}^J \{p_j \log \gamma\{\beta' \mathbf{z}_j\} - \gamma\{\beta' \mathbf{z}_j\} e_j\}, \quad (7)$$

où

$$e_j = \sum_{i=1}^N \int_{t_i^0}^{t_i^1} \lambda^*(u) du \quad (8)$$

est la contribution au nombre attendu de décès provenant de toutes les personnes-années d'observation dans l'état S_j , et p_j est le nombre total de décès dans cet état. En posant que $V_j(\beta) = \log \gamma\{\beta' \mathbf{z}_j\}$, nous obtenons l'estimation du maximum de vraisemblance $\hat{\beta}$ de β en tant que solution de l'équation de score

$$\frac{\partial \log L}{\partial \beta} = \frac{\partial \log L}{\partial V_j(\beta)} \frac{\partial V_j(\beta)}{\partial \beta} = \sum_{j=1}^J \frac{p_j}{e_j} \{p_j - \exp\{V_j(\beta)\} e_j\} = 0. \quad (9)$$

3. L'effet des erreurs de couplage sur les nombres observés et attendu de décès

Deux grands types d'erreurs peuvent se produire lors du couplage de fichiers de données dans le contexte du CIE (Fellegi et Sunter 1969). Un résultat faussement positif a lieu quand un membre de la cohorte encore en vie est

Les études par couplage d'enregistrements offrent plusieurs avantages par rapport aux études épidémiologiques classiques. L'utilisation des bases de données administratives existantes évite de devoir recueillir de nouvelles données pour les études sur la santé et permet d'obtenir des échantillons de grande taille moyennant assez peu d'efforts. Selon la nature des bases de données utilisées, le couplage d'enregistrements offre un moyen peu coûteux d'explorer de nombreuses associations éventuelles dans le cadre des études épidémiologiques. Le couplage d'enregistrements présente aussi certains inconvénients. Les chercheurs exercent généralement fort peu de contrôle sur l'information recueillie et le nombre de sujets perdus de vue lors des suivis peut être important. Les erreurs de couplage, qui sont le sujet du présent article, sont un autre inconvénient du couplage d'enregistrements. Inévitablement, certains enregistrements concordants ne seront pas couplés et certains enregistrements non concordants seront couplés incorrectement.

Assez peu de travaux ont été accomplis en vue de déterminer l'effet de ces erreurs de couplage sur les inférences statistiques. Neter, Maynes et Ramnathan (1965) ont utilisé un modèle de régression linéaire simple pour analyser l'effet des erreurs introduites durant le processus d'appariement. Selon leurs résultats, les erreurs de couplage font augmenter la variance résiduelle et introduisent un biais dans l'estimation de la pente de la droite de régression. Winkler et Scheuren (1991) établissent une expression du biais dû aux erreurs de couplage dans les estimations des coefficients de régression linéaire. Les progrès concernant l'estimation des taux d'erreurs de couplage réalisés par Belin et Rubin (1991) ont permis à Scheuren et Winkler (1993) de mettre en œuvre une méthode améliorée de correction du biais. L'application des méthodes de régression linéaire à l'analyse des fichiers de données appariées informatiquement est discutée plus en détail par Scheuren et Winkler (1997).

L'objet du présent article est d'étudier l'effet des erreurs de couplage sur les inférences statistiques dans les études-cohorte de la mortalité. À la section 2, nous décrivons les modèles de régression du risque relatif employés pour analyser les données provenant de ce genre d'études et nous élaborons des expressions pour les nombres observé et attendu de décès sur ces modèles. À la section 3, nous discutons de l'effet des erreurs de couplage sur les nombres observé et attendu de décès et de personnes-années à risque. À la section 4, nous analysons l'effet des erreurs de couplage sur les estimations des ratios standardisés de mortalité (RSM) et sur les paramètres de régression du risque relatif. Les deux types d'erreurs peuvent introduire un biais et une variabilité supplémentaire dans les estimations

conformément à des procédures bien établies en vue d'assurer le respect de la confidentialité des données (Singh, Feder, Dunteman et Yu 2001). Tous les fichiers couplés contenant des renseignements permettant d'identifier des individus restent sous la garde de Statistique Canada (Labossière 1986).

Des méthodes informatisées de couplage d'enregistrements ont été utilisées pour coupler des données sur l'exposition environnementale à celles de la Base canadienne de données sur la mortalité (BCDM). Par exemple, une étude a été entreprise pour étudier les liens éventuels entre les causes de décès chez plus de 326 000 exploitants agricoles au Canada et diverses variables sociodémographiques et d'exploitation agricole, particulièrement l'utilisation de pesticides (Jordan-Jimerson, Fair et Poliquin 1990). Cette étude comportait le couplage des données de la BCDM à celles du Recensement de la population de 1971 et du Recensement de l'agriculture de 1971. Une autre étude permanente de grande portée est fondée sur le Fichier dosimétrique national (FDN) du Canada (Ashmore et Grogan 1985; Ashmore et Davies 1989). Le FDN contient des renseignements remontant jusqu'à 1950 sur les expositions professionnelles aux rayonnement ionisants subies par plus de 400 000 Canadiens. Récemment, les enregistrements du FDN ont été couplés à ceux de la BCDM en vue d'étudier les associations entre la surmortalité due au cancer et l'exposition professionnelle à de faibles niveaux de rayonnements ionisants (Ashmore, Krewski et Zielinski 1997; Ashmore, Krewski, Zielinski, Jiang, Semenciv et Lévesque 1998). Plus récemment, les enregistrements du FDN ont été couplés à ceux de la Base canadienne de données sur l'incidence du cancer (Song, Zielinski, Ashmore, Jiang, Fair, Band et Lévesque 2001). La liste complète des autres études relatives à la santé fondées sur le couplage de données sur l'exposition à celles de la BCDM a été dressée par Fair (1989).

Le succès des études axées sur le couplage d'enregistrements dépend de la qualité des données couplées (Roos, Sooden et Jébamani 2001). À l'aide de données administratives longitudinales représentatives de la population, Roos et coll. ont examiné les problèmes de qualité de données dans les études sur l'état de santé et les soins de santé. Ardal et Ennis (2001) ont tenu compte des erreurs systématiques présentes dans les bases de données administratives intervenant dans l'analyse secondaire de l'information sur la santé. S'il est vrai que les études fondées sur le couplage d'enregistrements donnent de meilleurs résultats quand les données sont de haute qualité, les contraintes liées à la qualité des données sont compensées dans une certaine mesure par la grande taille des échantillons sur lesquels reposent de nombreuses bases de données administratives.

L'effet des erreurs de couplage d'enregistrements sur les estimations du risque dans les études-cohorte de mortalité

D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski et R. Mallick¹

Résumé

L'élaboration de la méthodologie de couplage informatisé d'enregistrements a facilité la réalisation d'études-cohorte de mortalité dans lesquelles les données sur l'exposition provenant d'une base de données sont couplées électroniquement à celles sur la mortalité provenant d'une autre base de données. Cependant, cette méthode donne lieu à des erreurs de couplage causées par l'appariement incorrect d'une personne figurant dans l'une des bases de données à une personne différente dans l'autre base de données. Dans le présent article, nous examinons l'effet des erreurs de couplage sur les estimations d'indicateurs épidémiologiques du risque, comme les ratios standardisés de mortalité et les paramètres des modèles de régression du risque relatif. Nous montrons que les effets sur les nombres observés et attendus de décès sont de sens opposé et que, par conséquent, ces indicateurs peuvent présenter un biais et une variabilité supplémentaire en présence d'erreurs de couplage.

Mots clés : Étude de cohorte; couplage informatisé d'enregistrements; erreurs de couplage; poids seuil de couplage; régression de Poisson; régression du risque relatif; ratio standardisé de mortalité.

1. Introduction

Ces dernières années, plusieurs études de cohorte historiques ont été réalisées en épidémiologie environnementale en se servant de bases de données administratives existantes comme sources d'information (Howe et Spasoff 1986; Carpenter et Fair 1990). En termes généraux, cette approche consiste à coupler des enregistrements de données sur l'exposition humaine à des risques environnementaux à des enregistrements de données sur l'état de santé, souvent au moyen de méthodes informatisées d'appariement d'enregistrements individuels provenant de bases de données différentes. Dans le cas d'une étude-cohorte de mortalité, le statut vital de chaque membre de la cohorte est déterminé par couplage aux enregistrements de décès des bases de données sur la mortalité tenues à jour par les organismes gouvernementaux. L'existence d'une surmortalité dans la cohorte comparativement à la population générale pourrait être due aux expositions subies par les membres de la cohorte.

En termes spécifiques, le couplage d'enregistrements est le processus consistant à regrouper deux ou plusieurs éléments d'information enregistrés distincts concernant une même entité (Bartlett, Krewski, Wang et Zielinski 1993). Les procédures de couplage informatisé d'enregistrements (CIB) sont devenues de plus en plus perfectionnées, grâce à

L'utilisation d'algorithmes complexes pour évaluer la probabilité que l'appariement de deux enregistrements soit correct (Hill 1988; Newcombe 1988). Statistique Canada a mis au point un système de CIB appelé CANLINK capable de coupler les enregistrements d'un même fichier, ainsi que ceux de deux fichiers distincts (Howe et Lindsay 1981; Smith et Silins 1981). Ce système attribue à chaque paire d'enregistrements un poids reflétant la probabilité qu'il s'agisse d'un appariement. Deux seuils sont fixés : les appariements potentiels dont le poids de couplage est supérieur au seuil supérieur sont considérés comme des coupages, tandis que les appariements potentiels dont le poids de couplage est inférieur au seuil inférieur sont considérés comme des non-couplages. Les cas d'appariement possible dont le poids est compris entre les seuils inférieur et supérieur sont résolus à l'aide de renseignements supplémentaires, lorsqu'ils sont disponibles. Sinon, on choisit un seuil unique pour faire la distinction entre les coupages et les non-couplages.

Lors de toute étude comportant un couplage d'enregistrements, des mesures strictes sont prises pour assurer la non-divulguation des enregistrements protégés aux termes de la *Loi sur la statistique*. Toutes les études qui nécessitent le couplage d'enregistrements faisant partie de bases de données protégées doivent être soumises à un processus d'examen et d'approbation rigoureux avant d'être exécutées

1. D. Krewski, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada K1N 6N5, School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6.
J. M. Zielinski, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; R. Mallick, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada, K1N 6N5, School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6.
Y. Wang, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; S. Bartlett, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; J. M. Zielinski, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; A. Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata, India; R. Mallick, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada, K1N 6N5, School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6.

- Matchware Technologies Inc. (1996). *AutoMatch: Generalized Record Linkage System User's Manual*. Silver Spring, MD: Matchware Technologies, Inc.
- Newcombe, H.B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford University Press, New York.
- Newcombe, H.B., Kennedy, J.M., Axford, S.J. et James, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 934-959.
- Newcombe, H.B., et Kennedy, J.M. (1962). Record linkage: Making maximum use of the discriminative power of identifying information. *Communications of the Association for Computing Machinery*, 5, 563-567.
- Oh, H.T., et Scheuren, F. (1980). Fiddling around with nonmatches. *Studies from Interagency Data Linkages Series*, Social Security Administration, Rapport No. 11.
- Scheuren, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 377-381.
- Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.
- Scheuren, F., et Winkler, W.E. (1997). Analyse de régression des fichiers de données appariées par ordinateur - Partie II. *Techniques d'enquête*, 23, 171-180.
- S-plus 2000 (1999). MathSoft, Inc. Data Analysis Products Division, Seattle, Washington.
- Tepping, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- Winglee, M., Valliant, R., Brick, J.M. et Machlin, S. (2000). Probability matching of medical events. *Journal of Economic and Social Measurement*, 26, 129-140.
- Winkler, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 829-834.
- Winkler, W.E. (1994). *Advanced Methods for Record Linkage*. Bureau of the Census Statistical Research Division, Statistical Research Report Series, RR 94/05.
- Winkler, W.E. (1995). *Matching and record linkage*. Dans *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christanson, M.J. Colledge et P.S. Kott) New York: John Wiley & Sons, Inc., 355-384.
- Agency for Healthcare Research et Quality (2001). MEP – Medical Expenditure Panel Survey. <<http://www.ahrq.gov/data/mepsix.htm>>.
- Armstrong, J.B., et Mayda, J.E. (1993). Estimation modeliste des taux d'erreur liés au couplage d'enregistrements. *Techniques d'enquête*, 19, 147-158.
- Barlett, S., Krewski, D., Wang, Y. et Zielinski, J.M. (1993). Evaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisés. *Techniques d'enquête*, 19, 3-13.
- Box, G.E.P., et Cox, D.R. (1964). An analysis of transformations (avec discussion). *Journal of the Royal Statistical Society, Series B*, 26, 206-252.
- Belin, T.R. (1993). Évaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle. *Techniques d'enquête*, 19, 15-33.
- Belin, T.R., et Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.
- Chambers, J.M., Cleveland, W.S., Kleiner, B. et Tukey, P. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.
- Fellegi, I.P. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Fellegi, I.P. (1997). Record linkage and public policy – A Dynamic Evolution. *Proceedings of the International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management and Budget*, Washington, DC.
- Gomati, S., Carter, R., Arter, A. et Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine*, 21, 1485-1496.
- Jaro, M.A. (1989). Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- Johnson, N.L., Kotz, S. et Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley & Sons, Inc.
- Lahiri, P., et Larsen, M.D. (2002). Regression analyses with linked data. (Manuscript d'ébauche).
- Larsen, M.D., et Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.

Bibliographie

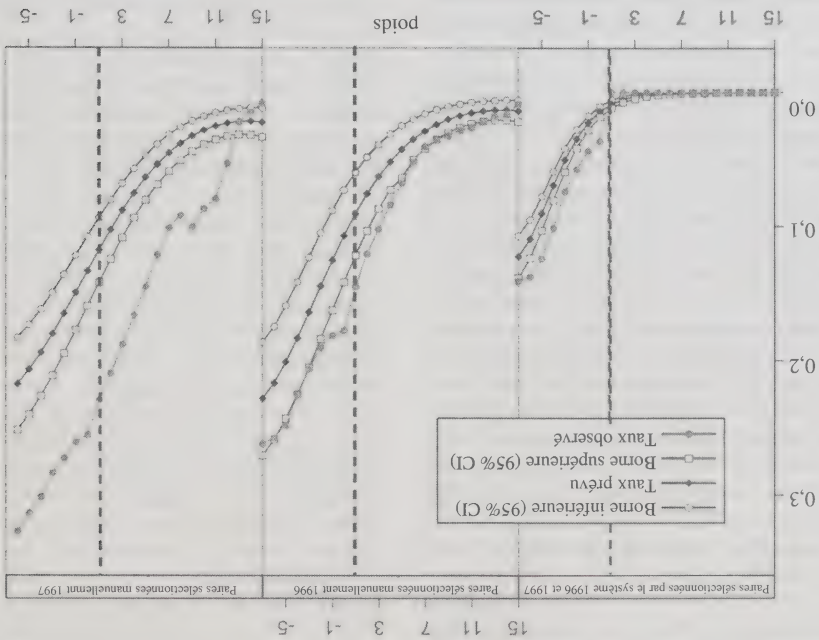


Figure 2. Estimations d'après le modèle de mélange de lois des taux de faux appariements selon le poids, échantillons d'apprentissage tirés de la MEPS de 1996 et de 1997 (une droite verticale est tracée au poids = 1, qui est le seuil).

Dans notre application, SimRate s'est avérée être un outil informatif et souple pour la détermination des seuils de sélection et l'estimation des taux d'erreurs. Étant donné un modèle multinomial ou d'autres modèles pour les variables d'appariement, la méthode SimRate fournit des estimations du taux d'erreurs que l'on obtiendrait par application répétée de l'algorithme d'appariement à un grand nombre de paires d'enregistrements candidates. Elle s'avère également souple en ce qui concerne le choix des ensembles de paires à comparer pour calculer les taux.

Bien que notre application nous ait permis de réaliser nos objectifs d'estimations des taux d'appariement et des taux d'erreurs pour la MEPS, une étude plus approfondie pourrait être réalisée avant l'étape de l'analyse ou durant travaux dans le contexte de la présente étude de cas, mais celle-ci. Faute d'espace, nous ne pouvons élaborer ces nous pourrions mentionner deux approches générales. En premier lieu, il est possible de répondre les résultats finals et de les corriger pour les faux non-appariements, en traitant ceux-ci d'une façon analogue à la non-réponse unitaire

Remerciements

L'étude fondamentale du couplage d'enregistrements présentée ici a été réalisée en vertu des contrats 290-99-0002 et 290-94-2002 parrainés par l'Agency for Healthcare Research and Quality et le National Center for Health Statistics. Les auteurs remercient Steven B. Cohen, Steven Machlin et Joel Cohen, de l'Agency for Healthcare Research and Quality, de leurs commentaires à diverses étapes de cette étude, et Thomas Bellin, pour ses suggestions concernant une version antérieure.

(par exemple, comme dans Oh et Scheuren 1980). Pour traiter les appariements incorrects, les idées proposées dans Scheuren et Winkler (1993 et 1997) et dans Lahiri et Larsen (2002) vaudraient peut-être la peine d'être consultées. La question de savoir si ces étapes supplémentaires sont nécessaires dépend, évidemment, de l'utilisation finale prévue des données couplées.

qualité des appartements déclarés, comme nous l'avons constaté dans le cas de la MEPS.

La méthode d'estimation fondée sur les courbes de distribution des poids a l'avantage de permettre de choisir un seuil de sélection pour atteindre le niveau acceptable d'erreurs de couplage. Par exemple, la figure 1 montre l'échantillon d'apprentissage et les courbes de distribution des poids simulées au moyen de SimRate d'après les fichiers d'appariement de la MEPS de 1996. Une droite verticale est tracée au poids seuil de sélection $w = 1$; les niveaux d'erreur pour la MEPS de 1996 (présentés au tableau 3) ont alors été estimés par le pourcentage cumulé au niveau seuil. En déplaçant ce seuil, on peut tenter de réduire au minimum l'erreur totale de couplage en choisissant un seuil au point d'intersection des courbes M et U . Dans la présente étude de cas, les seuils optimaux suggérés par les deux ensembles de courbes de distribution des poids sont assez cohérents. Nous avons inclus une échelle du rapport de vraisemblance dans la figure pour donner une interprétation grossière de la vraisemblance du poids d'appariement. Par exemple, pour le poids d'appariement $w = 1$, le rapport de vraisemblance est égal à 2. Autrement dit, pour les enregistrements dont le poids

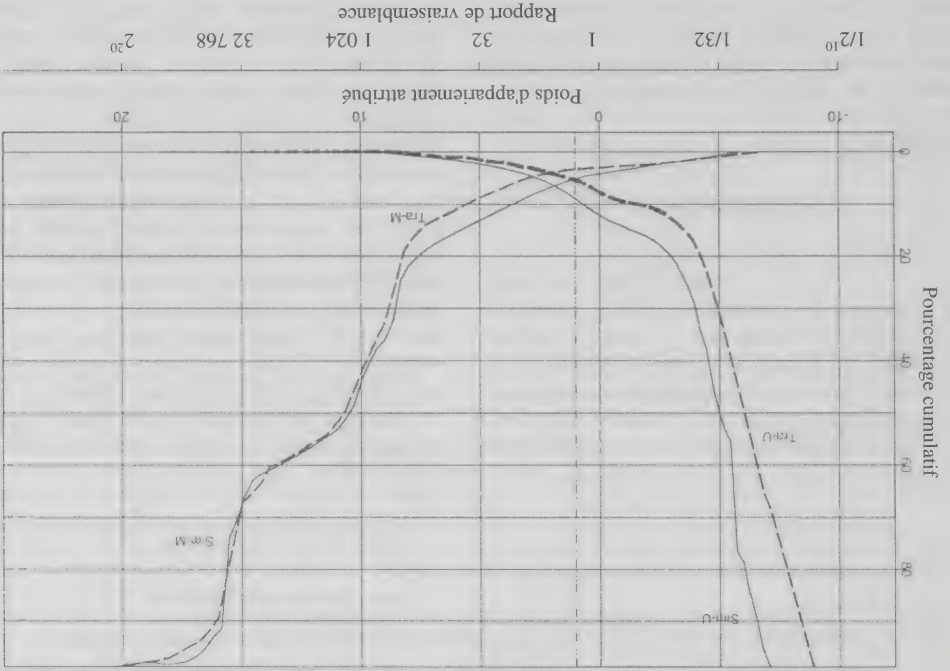


Figure 1. Courbes des poids pour la MEPS de 1996 d'après les méthodes SimRate et d'échantillons d'apprentissage; la droite de référence verticale en pointillé montre la valeur seuil de 1.

d'appariement est égal ou supérieur à $w = 1$, la vraisemblance relative qu'il s'agisse d'un vrai appariement est au moins de 2 à 1.

En ce qui concerne la qualité des paires couplées, la figure 2 montre les distributions des estimations des taux de faux appariements calculés d'après le modèle de mélange de lois. Elle donne le taux de faux appariements estimé d'après le modèle, les bornes supérieure et inférieure de l'intervalle de confiance à 95 % des estimations du taux d'erreurs et les taux réels observés. Le premier panneau montre les estimations lorsqu'on traite les paires couplées sélectionnées par le système informatique comme étant des appariements vrais. Les deuxième et troisième panneaux montrent les estimations produites d'après les échantillons d'apprentissage tirés de la MEPS de 1996 et de la MEPS de 1997. La différence entre les deuxième et troisième panneaux montre le manque d'uniformité de la sélection manuelle par divers examinateurs dans notre application. Dans aucun des trois panneaux l'intervalle de confiance à 95 % des estimations d'après le modèle ne couvre les valeurs réelles observées. Idéalement, on devrait utiliser à la fois la figure 1 et la figure 2 pour orienter le choix des seuils de sélection.

niveau d'erreurs ne sera pas nul, si bien que l'intervalle de confiance pour le modèle de mélange de lois n'est pas nécessairement erroné.

Tableau 4
Estimations de l'erreur de couplage par la méthode du modèle de mélange de lois

MEPS de 1996	Pourcentage de faux appariements		Taux observé
	Taux Borne inférieure*	Taux Borne supérieure*	
Appariement manuel	9,1	6,0	12,2
Appariement par le système	0,9	0,6	1,2
			0,0

* Les bornes inférieure et supérieure sont celles de l'intervalle de confiance à 95 % du taux d'erreur prévu.

Nous avons produit des estimations globales des paramètres en utilisant les échantillons d'appariements sélectionnés manuellement ainsi que par le système pour les MEPS de 1996 et de 1997, afin de créer quatre ensembles de données d'entrée pour produire des estimations globales pour modéliser l'erreur de couplage pour la MEPS de 1998. Cela a été possible, parce que les données sont restées semblables et que les paires d'enregistrement ont été sélectionnées en appliquant les mêmes règles d'appariement pour les trois années. La seule différence était que nous n'avons pas procédé à un examen manuel pour la MEPS de 1998 et que nous n'avons pas pu utiliser la procédure de Box-Cox pour l'estimation globale des paramètres pour les paires vraies et fausses). Pour cette application, nous avons utilisé une méthode bootstrap dans la procédure de calage de Belin-Rubin afin de nous appuyer sur plusieurs ensembles de paramètres de façon à refléter les incertitudes de l'estimation. Cependant, cette application n'a pas convergé après 150 itérations de la procédure d'estimation. Nous avons seulement pu conclure que les échantillons d'appariements antérieurs ne pouvaient être généralisés et fournir des estimations du taux d'erreurs pour des applications de couplage répétées.

8. Conclusion et incidences analytiques

La sélection d'un seuil et l'estimation de l'erreur de couplage représente un processus itératif comprenant des cycles répétés d'observation, d'estimation et de modélisation. Notre étude s'appuie sur des approches de modélisation pour estimer les erreurs de couplage et évaluer le pouvoir prédictif du système de couplage. Les deux méthodes fournissent des renseignements valables pour déterminer la sélection des couplages et pour évaluer la

où w_j est le poids d'appariement pour la paire r , w est la moyenne géométrique des poids w_j , et γ est un paramètre qui dépend du fait que la paire appartient à l'ensemble de paires appariées ou non appariées.

Pour que la méthode de mélange de lois soit efficace, les poids transformés doivent suivre une loi approximativement normale. La distribution des poids non observés obtenus au moyen de nos données indiquait une bimodalité et pratiquement aucun chevauchement entre les poids d'appariement des paires appariées et non appariées. (Belin-Rubin 1995 ont également observé une bimodalité). Par exemple, l'application de leur procédure de transformation aux paires sélectionnées par le système pour la MEPS de 1996 a donné les estimations des paramètres $\underline{w} = 58,7$ et $\gamma = 1,15$ pour les paires vraiment appariées et $\underline{w} = 113,1$ et $\gamma = 0,48$ pour les paires faussement appariées. Cependant, le rapprochement des poids transformés vers la loi normale était assez faible. Puisque les poids d'appariement sont égaux au logarithme d'un produit, c'est-à-dire à la somme des logarithmes des termes de ce produit, nous pourrions espérer que les poids suivent une loi normale si la somme compte un grand nombre de composantes. Cependant, nous ne disposons que de cinq zones pour procéder à l'appariement. Le petit nombre de zones pourrait expliquer, en partie, l'écart de nos données transformées par rapport à la loi normale.

Le tableau 4 donne les résultats de l'application du modèle de mélange de lois de Belin-Rubin aux données de la MEPS de 1996. Il contient les taux de faux appariements estimés d'après le modèle, l'intervalle de confiance à 95 % du taux estimé et le taux de faux appariements réels observés au poids seuil de 1. En considérant les paires déterminées par examen manuel comme étant des appariements vrais, une estimation fondée sur le modèle du taux prévu d'appariements faux au seuil de $w = 1$ était de 9,1 %, avec un intervalle de confiance à 95 % variant de 6,0 à 12,2. Par contre, le taux réel observé de faux appariements était de 14,5 %, valeur plus élevée que la borne supérieure de l'intervalle de confiance à 95 %. Il convient de souligner qu'il s'agit des taux de la forme n_2/n_1 , du tableau 1, qui ne sont pas les mêmes que les taux estimés par la méthode Simkate et par la méthode des courbes de poids.

Puisque l'examen manuel n'est pas forcément toujours exact, une option, aux fins d'évaluation, consiste à traiter les paires coupées sélectionnées par le système informatique comme étant les paires vraiment appariées et à les utiliser pour la modélisation. Sous cette hypothèse, l'estimation fondée sur le modèle du taux d'erreurs prévu est de 0,9 et l'intervalle de confiance à 95 % varie de 0,6 à 1,2. Le taux réel observé dans ce cas, c'est-à-dire 0 %, est un résultat hypothétique où les paires coupées sélectionnées par l'ordonateur sont traitées comme étant correctes. En réalité, le

Lorsqu'on dispose d'estimations paramétriques globales des paramètres transformés et du ratio des variances des deux distributions, on peut les appliquer à des données similaires pour l'estimation. Puisque le couplage des enregistrements de la MEPS est réalisé annuellement, des estimations globales calculées d'après des échantillons d'apprentissage antérieurs pourraient, en théorie, être appliquées à l'estimation de l'erreur de couplage lors d'années ultérieures si l'on ne dispose pas d'échantillons pour l'examen manuel.

Le deuxième avantage est que le modèle de mélange de loirs peut s'appuyer sur plusieurs ensembles d'estimations de paramètres provenant de divers échantillons d'apprentissage et refléter les variations. Cette caractéristique est particulièrement séduisante dans le cas de la MEPS, parce que l'examen manuel est un processus complexe, qui n'est pas forcément toujours exact. Donc, une autre solution consiste à considérer les paires sélectionnées par le système informatique comme étant des appariements vrais et à les utiliser pour produire un ensemble d'estimations des paramètres de rechange. Ce processus peut également être répété en utilisant les échantillons d'apprentissage provenant de plus d'une année.

Dans notre application de l'approche de Belin-Rubin, nous avons utilisé les mêmes échantillons d'apprentissage provenant de la MEPS de 1996, ainsi qu'un deuxième échantillon d'apprentissage de même taille provenant de l'enquête de 1997. Suivant l'exemple de Belin-Rubin, nous avons appliqué la méthode du modèle de mélange de loirs en utilisant des paires reconnues manuellement comme étant des appariements vrais ou faux produites par un système de appariement binivoque (un vers un) (il convient de souligner que ce genre de système produit assez peu de paires représentant de faux appariements pour l'estimation). Nous avons calculé les estimations fondées sur le modèle pour la MEPS de 1996 et la MEPS de 1997 en supposant que la sélection manuelle était correcte et, pour tester le comportement du modèle, nous avons calculé un deuxième ensemble d'estimations en supposant que les paires sélectionnées par le système informatique comme étant des appariements étaient les paires correctes.

L'application de la méthode comportait deux procédures, à savoir, la procédure de Box et Cox (1964) pour l'estimation globale des paramètres et la procédure de calage (Belin et Rubin 1995) pour ajuster un modèle de mélange de loirs en vue d'estimer le taux d'erreurs. Avant d'appliquer la méthode de Box-Cox, nous avons rééchantonné les poids entre 1 et 1 000. La transformation de Box-Cox discutée par Belin et Rubin (1995) était

$$\psi(w_i) = \frac{w_i}{w_i - 1}$$

facile à répéter chaque année. Par contre, la répétition des courbes de poids établies manuellement dépendait en partie de l'examen manuel et nous ne disposons que d'un seul échantillon d'apprentissage fiable, c'est-à-dire celui obtenu pour 1996. Il convient de souligner que les paires couplées utilisées dans SimRake génèrent naturellement un certain pourcentage de résultats faussement positifs et de résultats faussement négatifs, c'est-à-dire certaines paires considérées incorrectement comme étant appariées, d'une part, et non appariées, d'autre part. Donc, les probabilités m_{ij} calculées de cette façon pour les zones mentionnées peuvent comporter une erreur. Il aurait été préférable d'estimer les probabilités m à partir d'un ensemble « vrai » pour lequel nous étions certains que tous les appariements étaient corrects. Cependant, les ensembles d'apprentissage appariés manuellement que nous avons pu produire étaient trop petits pour donner des estimations stables pour toutes les catégories de détaillées d'appariement et, qui plus est, la sélection manuelle est également imparfaite. Cette différence pourrait expliquer, du moins partiellement, les estimations légèrement plus élevées du taux d'erreurs global produites par SimRake comparativement à celles obtenues d'après les courbes de poids établies pour l'échantillon d'apprentissage.

Tableau 3
Méthodes des courbes de poids pour estimer les taux d'erreurs de couplage au poids seuil de 1, MEPS 1996 à 1998

Méthode	Taux d'erreurs	1996	1997	1998
Courbes de simulation	Faussement négatifs	5,2	6,5	5,8
SimRake	Faussement positifs	9,0	6,9	7,6
Courbe de l'échantillon	Faussement négatifs*	3,3	3,3	3,3
d'apprentissage	Faussement positifs**	5,5	6,4	5,7

* Les estimations établies d'après la courbe Tra-M de 1996 ont été utilisées pour les trois années.
** Les estimations d'après la courbe Tra-U de 1996 ont été produites au moyen d'échantillons de 500 enregistrements provenant de chaque fichier d'appariement et d'un total de 250 000 paires non appariées. Les estimations pour 1997 et pour 1998 ont été produites au moyen d'autres courbes Tra-U fondées sur des échantillons de 1 000 enregistrements provenant de chaque fichier d'appariement et un total de 1 000 000 de paires non appariées.

7. Application des modèles de mélange de loirs à la MEPS

Dans leur approche par modélisation d'un mélange de loirs, Belin et Rubin (1995) considéraient la distribution des poids d'appariement observés à partir d'un système automatisé de couplage d'enregistrements comme étant un mélange de poids pour les vrais appariements et les faux appariements. En principe, la méthode du modèle de mélange de loirs possède deux caractéristiques intéressantes qui conviennent pour la MEPS. En premier lieu, l'application répétée de la méthode peut se faire efficacement.

de l'événement, 9 catégories pour la durée de l'hospitalisation, 27 catégories pour l'intervention médicale et 3 catégories pour le problème de santé, ainsi que pour les frais globaux. Par exemple, pour la catégorie de résultats Concordance exacte de la date de l'événement, l'estimation de m_{vi} était 0,69, ce qui signifie que la concordance de la date de l'événement était exacte pour 69 % des paires couplées. L'estimation de u_{vi} pour cette catégorie de résultats était 0,003, montrant que 0,3 % seulement des paires non couplées présentaient une concordance pour cette zone. Le poids d'appariement pour la concordance exacte de la date de l'événement était 8,52 et celui pour la non-concordance complète (différence de plus de deux semaines et jour de la semaine différent) était -6,64 (voir Winglee et coll. 2000 pour les poids d'appariement selon la zone d'appariement et la catégorie de résultats).

Tableau 2

Estimations des probabilités multinomiales pour les paires appariées (m_{vi}) et pour les paires non appariées (u_{vi}), et poids d'appariement (w_{vi}) pour la zone d'appariement Date de

Règle d'appariement pour la date de l'événement		l'événement	
m_{vi}	u_{vi}	m_{vi}	w_{vi}
0,031	0,046	Appariement exact	0,003
0,068	0,006	Décalage de +/- 1 jour	0,547
0,023	0,005	Décalage de +/- 3 jours	0,034
0,014	0,005	Décalage de +/- 5 jours	0,034
0,030	0,006	Décalage de +/- 7 jours	0,034
0,014	0,006	Appariement du jour de la semaine	0,034
0,003	0,047	Non-concordance	0,034

Pour notre étude de cas, nous avons choisi des zones d'appariement qui étaient approximativement indépendantes. Ainsi, nous n'avons observé aucune association fonctionnelle entre la date de l'événement médical et d'autres zones d'appariement, comme celles du problème médical et de la durée de l'hospitalisation. Pour des zones comme celles des indicateurs d'intervention chirurgicale, d'examen radiologique et d'analyses biologiques, nous avons utilisé des tests du chi-carré et constaté une certaine dépendance entre l'intervention chirurgicale et l'examen radiographique concurrents. Pour contourner cette situation, nous avons estimé des probabilités conjointes et spécifié des règles d'appariement en vue de traiter ces indicateurs d'intervention comme une zone d'appariement unique (voir la section 4 plus haut). Par conséquent, nous avons pu appliquer la loi multinomiale indépendante pour la simulation. Le tableau 3 donne les résultats de l'estimation de l'erreur de couplage par la méthode SimKate et par celle des courbes d'apprentissage au poids seuil de $w = 1$ pour les MEPS de 1996, 1997 et 1998. La méthode SimKate a été

les courbes de poids. Pour générer les distributions de poids « Sim-M », nous avons estimé les probabilités m_{vi} d'après des paires couplées déterminées au moyen d'un algorithme d'appariement unique. Nous avons utilisé le système de couplage « règle » pour sélectionner des paires appariées d'après les fichiers annuels d'appariement de 1996 et nous avons utilisé la proportion de paires entrant dans la catégorie i de la zone v comme estimation m_{vi} de la probabilité m_{vi} .

estimation u_{vi} de la probabilité u_{vi} .

Pour une paire appariée simulée, nous avons généré une réalisation de la variable aléatoire multinomiale y_{vi} pour chaque zone d'appariement. Par exemple, nous avons généré une configuration telle que (concordance pour la date de l'événement, concordance pour la durée de l'hospitalisation, concordance pour la gamme de codes de problème médical, concordance conjointe selon le type d'intervention, et concordance pour une valeur spécifique d'un indicateur de frais globaux) en utilisant les probabilités d'appariement m_{vi} pour chaque catégorie de résultats. De la même façon, pour chaque paire non appariée, nous avons généré une réalisation d'une catégorie pour chacune des cinq zones en utilisant les probabilités d'appariement incorrectes u_{vi} mentionnées plus haut.

Pour une réalisation donnée y_{vi} , nous avons calculé un poids w_{vi} pour la paire en additionnant les poids pour les catégories générées aléatoirement dans lesquelles se classait la paire. Les poids réels utilisés dans notre simulation étaient des poids corrigés que nous avons spécifiés, plutôt que des poids définis directement par le logiciel d'appariement (voir Winglee et coll. 2000). Donc, nous avons simulé la façon dont l'appariement serait effectivement mis en œuvre. Pour cela, nous avons calculé le poids d'appariement pour les ensembles de 10 000 paires appariées et de 10 000 paires non appariées, et nous avons représenté graphiquement les fonctions simulées des poids d'appariement.

Le tableau 2 donne des exemples de catégories de concordance partielle utilisées pour l'appariement d'après la date de l'événement et les estimations de m_{vi} , u_{vi} et w_{vi} utilisées dans la simulation SimKate. Nous avons défini, en tout, 19 catégories de résultats pour l'appariement de la date

utiliser ces probabilités dans la simulation. Des données peu nombreuses limiteront naturellement le nombre de cellules pour lesquelles la méthode est applicable. Cependant, en présence de données peu nombreuses, la pénalité en cas d'échec du modèle doit être faible.

5. Couplage des enregistrements des événements médicaux de la MEPS

Le couplage des enregistrements des événements médicaux de la MEPS a été réalisé en utilisant cinq zones d'identification, à savoir la date de l'événement (année, mois, jour et jour de la semaine), les codes de problème médical, les codes d'intervention, le code de frais globaux et la durée (nombre de jours) de l'hospitalisation. Ces zones sont décrites plus en détail dans Wingée, Vaillant, Brick et Machlin (2000). Nous nous sommes servis d'un échantillon d'apprentissage tiré de la MEPS de 1996 pour établir les règles d'appariement et les catégories de résultats, ainsi que pour estimer les probabilités de concordances pour chaque catégorie, en tenant compte des concordances partielles et des valeurs particulières. Nous avons répété les mêmes règles d'appariement aux paramètres d'appariement.

Pour l'ensemble d'apprentissage, nous avons utilisé le système de couplage Automatch (Matchware 1996) et l'algorithme d'appariement unique pour sélectionner les paires couplées. L'appariement « unique » consiste à coupler de façon optimale un enregistrement du fichier A à un seul enregistrement du fichier B (Jaro 1989). En outre, nous avons utilisé l'algorithme d'appariement multi-multivoque (plusieurs vers plusieurs) pour générer un échantillon aléatoire de paires non couplées en vue de faciliter l'estimation de l'erreur de couplage. Cependant, les méthodes d'estimation des taux d'erreurs, qui sont décrites plus loin, s'appliquent à tout logiciel qui met en œuvre des méthodes de couplage fondées sur des poids d'appariement. Elles ne sont pas particulières à Automatch.

Afin de déterminer le seuil de sélection pour la MEPS, nous avons fait un compromis entre l'obtention d'un taux élevé d'appariement réels et la limitation des erreurs de couplage par appariement incorrect. Un poids seuil élevé réduirait au minimum le nombre de résultats faussement positifs (appariements incorrects), au prix d'une réduction du taux d'appariements corrects et d'une perte de données précieuses recueillies auprès des prestataires de soins médicaux. Par ailleurs, un seuil faible augmenterait le nombre de résultats faussement positifs et pourrait avoir sur la répartition des données sur les dépenses un effet dont on ne pourrait venir en bout qu'en recourant à des techniques analytiques spéciales et, même alors, avec incertitude seulement. Puisque les deux sources de données avaient fait

des déclarations sur manifestement les mêmes événements médicaux pour les mêmes personnes au cours de la même période, notre stratégie a été de maintenir un taux d'appariement raisonnablement élevé et de procéder à un examen manuel d'un nombre limité de paires couplées douteuses après sélection afin d'évaluer l'effet analytique de leur acceptation erronée. Basé sur cette décision, le taux moyen d'appariement réel pour les fichiers annuels d'événements médicaux de la MEPS était d'environ 85 %.

La courbe *M* pour l'échantillon d'apprentissage tiré de la MEPS de 1996, annotée courbe « *Tra-M* » a été produite en appliquant les poids d'appariement aux paires qui étaient des appariements « réels » pour un échantillon aléatoire de 500 personnes ayant participé à la MEPS de 1996. Pour ces 2 507 événements déclarés par les répondants des ménages médicaux. Des gestionnaires de données chevronnés ont passé les événements en revue et ont sélectionné 1 501 paires. Nous avons considéré ces dernières comme étant des appariements vrais dans la présente évaluation. Nous avons attribué aux paires appariées manuellement les poids déterminés d'après notre spécification d'appariement pour générer une fonction de distribution cumulative.

Nous avons produit la courbe *U* pour l'échantillon d'apprentissage de 1996, annotée courbe « *Tra-U* », au moyen d'un échantillon aléatoire de paires constituant des non-appariements. Nous avons appliqué une méthode d'échantillonnage aléatoire simple avec remise pour sélectionner 500 événements dans chacun des fichiers d'appariement et un algorithme d'appariement multi-multivoque (plusieurs vers plusieurs) pour générer les 250 000 paires d'événements possibles. Pour ces ensembles de paires sélectionnées au hasard, les chances qu'il existe une paire correctement appariée sont négligeables; par conséquent, nous avons considéré l'ensemble complet comme formé de paires incorrectement appariées. Nous avons appliqué les poids d'appariement obtenus selon notre spécification d'appariement et tracé la courbe « *Tra-U* » égale à 1 moins la distribution cumulative des poids de ces paires. La figure 1 de la section 8 montre les courbes *Tra-M* et *Tra-U* pour la MEPS de 1996. Ces courbes ont été lissées au moyen d'une fonction *lowess* non paramétrique (Chamber, Cleveland, Kleiner et Tukey 1983) en *S-PLUS* (2000 (1999)).

6. Application de Simkate à la MEPS

La méthode Simkate d'établissement des distributions des poids consiste à appliquer des méthodes de simulation de Monte Carlo pour générer des ensembles distincts de 10 000 paires appariées et non appariées simulées pour créer

Voir la section 6 sur les poids d'appariement utilisés pour la

simulation.

La distribution cumulative de ces poids pour les paires

appariées simulées est alors portée en graphique pour

produire la courbe « Sim- M ». De la même façon, la

distribution cumulative inverse des paires non appariées est

représentée graphiquement pour produire la courbe

« Sim- U » (voir la figure 1, à la section 8, pour un exemple

de toutes les courbes de simulation utilisées dans l'étude).

La proportion simulée de paires appariées dont les poids

sont inférieurs au seuil est l'estimation du taux de résultats

faussement négatifs. La proportion simulée de paires non

appariées dont le poids est supérieur au seuil représente

l'estimation du taux de résultats faussement positifs.

Cette approche requiert l'estimation empirique des

distributions des variables d'appariement tant pour les paires

réellement appariées que pour les paires réellement non

appariées. Même si l'algorithme de détermination des poids

repose sur l'hypothèse d'indépendance des variables

d'appariement, les données réelles peuvent témoigner d'une

dépendance. À condition de pouvoir générer des paires

artificielles qui suivent raisonnablement la loi observée des

données (en intégrant toute dépendance), alors cette

méthode devrait produire des estimations appropriées des

taux d'erreurs.

rentre. Le poids d'appariement w_i d'une paire d'enregistre-

ments est habituellement estimé comme suit.

$$w_i = \log_2 \left[\frac{\prod_{v=1}^V \prod_{c_v=1}^{C_v} m_{i,c_v}^{w_{i,c_v}}}{\prod_{v=1}^V \prod_{c_v=1}^{C_v} n_{i,c_v}^{w_{i,c_v}}} \right]$$

rentre la paire r . Pour chaque zone, l'une des valeurs de y^{r,i_1}

sera 1 et les autres, 0.

L'hypothèse théorique qui sous-tend l'approche SimKate

est celle selon laquelle y^{r,i_1} suit une loi multinomiale si la

paire r est un appariement réel et une loi multinomiale

différente si la paire est un non-appariement. Nous pouvons

alors modéliser les vecteurs y^{r,i_1} par une loi multinomiale de

paramètre $\mathbf{m}^{r,i_1} = (m_{r,i_1}^{c_1}, \dots, m_{r,i_1}^{c_{C_v}})$ si la paire est un appari-

ment vrai et de paramètre $\mathbf{n}^{r,i_1} = (n_{r,i_1}^{c_1}, \dots, n_{r,i_1}^{c_{C_v}})$ si elle est un

non-appariement. Alors, la probabilité $m_{r,i_1}^{c_v} = P$

(concordance de la catégorie i de la zone v dans la paire

$r \in M$) est la probabilité conditionnelle de concordance

pour la catégorie i de la zone v , sachant que la paire

d'enregistrements r est comprise dans l'ensemble M de

paires réellement appariées. Par contre, la probabilité

$n_{r,i_1}^{c_v} = P$ (concordance de la catégorie i de la zone v dans la

paire $r \in U$) est la probabilité conditionnelle de con-

cordance pour la catégorie i de la zone v , sachant que la

paire d'enregistrements r est comprise dans l'ensemble U de

paires réellement non appariées. En supposant que les

variables d'appariement, $v = 1, \dots, V$, sont indépendantes,

nous pouvons spécifier la probabilité conjointe de

$y^r = (y^{r,i_1}, \dots, y^{r,i_V})$ si une paire r est un appariement vrai

$$P(y^r | r \in M) = \prod_{v=1}^V \prod_{c_v=1}^{C_v} m_{y^{r,i_v}}^{w_{i_v}^{c_v}}.$$

sous la forme

La probabilité correspondante de la même configuration de

données, si la paire est réellement un non-appariement, est

$$P(y^r | r \in U) = \prod_{v=1}^V \prod_{c_v=1}^{C_v} n_{y^{r,i_v}}^{w_{i_v}^{c_v}}.$$

SimKate utilise des méthodes de simulation de Monte Carlo pour générer un grand nombre de réalisations de paires appariées et de paires non appariées en se fondant sur des estimations des probabilités $m_{i_1}^{c_v}$ et $n_{i_1}^{c_v}$. Pour chaque paire simulée, l'application calcule un poids d'appariement w_i , qui est appliqué à une configuration particulière de données. Pour une réalisation donnée y^r , un poids w_r est calculé pour la paire en additionnant les poids pour les catégories générées aléatoirement dans lesquelles la paire

L'ensemble de paires d'enregistrements appariées ou de paires d'enregistrements non appariées est multimodale. Une autre exigence essentielle est de disposer d'un ensemble de données d'appariement dont les caractéristiques sont semblables à celles qu'il faudra appariées. Faute de posséder un bon ensemble de données d'appariement, l'estimation des paramètres d'entrée du modèle de mélange de lois pourrait être médiocre, ce qui aurait une incidence sur les taux d'erreurs estimés finaux. Dans le cas de notre application, en utilisant des données annuelles sur des événements médicaux répétés sur trois années, les paramètres n'étaient pas stables au cours du temps. Cette instabilité nous a obligés à utiliser un ensemble d'appariements pour chaque année, ce qui rend l'approche de Bellin-Rubin peu pratique pour notre application, en raison du coût et du temps requis.

L'approche par simulation, SimKate, offre, comme la modélisation de mélange de lois, la capacité d'examiner divers seuils, ce qui permet à l'utilisateur de surveiller à la fois la sensibilité et la spécificité de la règle de décision en vue de sélectionner les paires appariées. À condition de pouvoir modéliser raisonnablement le processus utilisé pour établir les poids d'appariement, il est possible d'appliquer des méthodes personnalisées d'attribution de poids, telles que celles utilisées pour la présente étude de cas. La méthode requiert la production de paires d'enregistrements en se fondant sur la distribution des caractéristiques des ensembles de paires appariées et non appariées. Un certain effort doit être déployé pour générer raisonnablement les populations de paires. Dans le cadre de nos travaux, nous avons réussi à générer ces populations au moyen de modèles multinomiaux.

3. Poids seuils et estimation de l'erreur de couplage

Plusieurs méthodes sont décrites dans la littérature pour sélectionner les appariements vrais et pour estimer les erreurs de couplage (par exemple, Bartlett, Krewski, Wang et Zielinski 1993; Armstrong et Mayda 1993; Bellin 1993; Bellin et Rubin 1995; Winkler 1992, 1995). Consulter Fellegi (1997) pour une vue d'ensemble de l'évolution du couplage d'enregistrements, Tepping (1968) et Larsen et Rubin (2001), pour d'autres méthodes de couplage et Scheuren (1983), pour une méthode de capture-recapture pour estimer l'erreur d'omission.

La comparaison des estimations obtenues selon les diverses approches se complique du fait que chacune a tendance à se concentrer sur des composantes différentes de l'erreur. En fait, les méthodes exposées dans la littérature sur le couplage d'enregistrements pour calculer des taux

Nous utilisons deux options fondées sur un modèle pour estimer l'erreur de couplage. L'une repose sur la simulation pour obtenir une distribution des poids pour divers niveaux de concordance. Cette technique, appelée SimKate, permet la génération des distributions des poids pour les paires d'enregistrements appariées et non appariées. Partant de ces distributions, SimKate peut alors fournir des estimations des taux d'erreurs de couplage pour divers seuils de sélection. Ces taux d'erreurs peuvent ensuite servir de guide pour apporter des corrections et pour évaluer le succès de l'opération de couplage. Nous comparons la méthode SimKate à une deuxième méthode de modélisation élaborée par Bellin et Rubin (1995). Comme nous espérons le montrer, ces approches ont toutes deux leur place; chacune possède des points forts, comme l'illustre les comparaisons.

2. Modèles de mélange de lois et approche SimKate

La méthode d'estimation de l'erreur de couplage par modélisation de mélange de lois présentée dans Bellin et Rubin (1995) possède plusieurs caractéristiques intéressantes. Elle est souple en ce sens que le processus d'établissement des poids n'a pas à être considéré directement. Par conséquent, cette méthode est applicable à de nombreuses méthodes de création des poids. Lorsqu'un modèle est spécifié, on peut examiner les taux d'erreurs pour un continuum de valeurs seuils potentielles et construire des bandes de confiance pour surveiller la précision des estimations de l'erreur (voir la section 7).

Toutefois, les modèles de mélange de lois ont leurs limites. La méthode fournit un taux particulier d'erreurs, à savoir la proportion d'enregistrements couples qui représente effectivement des non-appariements, mais il est impossible d'estimer les taux de résultats faussement positifs ou faussement négatifs, puisqu'on ne tient pas compte des paires non couplées. Le taux d'erreur estimé est un taux conditionnel qui dépend de l'ensemble de paires d'enregistrements qui ont été couplées. De surcroît, les paramètres du modèle peuvent être difficiles à estimer si l'on ne peut isoler les distributions des poids pour les ensembles de paires d'enregistrements appariées et non appariées (voir Winkler 1994).

L'une des hypothèses importantes qui sous-tend l'approche de Bellin-Rubin est qu'il est possible de transformer les distributions des poids dans les ensembles de paires d'enregistrements appariées et non appariées (voir Winkler 1994).

Une étude de cas en couplage d'enregistrements

M. Winglee, R. Valliant et F. Schuren¹

Résumé

Le couplage d'enregistrements est un processus qui consiste à apparier des enregistrements provenant de deux fichiers en essayant de sélectionner les paires dont les deux enregistrements appartiennent à une même entité. La démarche fondamentale consiste à utiliser un poids d'appariement pour mesurer la probabilité qu'un appariement soit correct et une règle de décision pour décider si une paire d'enregistrements constitue un « vrai » ou un « faux » appariement. Les seuils de poids utilisés pour déterminer si une paire d'enregistrements représente un appariement ou un non-appariement dépend du niveau de contrôle souhaité sur les erreurs de couplage. Les méthodes appliquées à l'heure actuelle pour déterminer les seuils de sélection et estimer les erreurs de couplage peuvent donner des résultats divergents, selon le type d'erreur de couplage et la méthode de couplage. L'article décrit une étude de cas reposant sur les méthodes existantes de couplage pour former les paires d'enregistrements, mais sur une nouvelle approche de simulation (Simkare) pour déterminer les seuils de sélection et estimer les erreurs de couplage. Simkare s'appuie sur la distribution observée des données dans les paires appariées et non appariées afin de générer un grand ensemble simulé de paires d'enregistrements, d'attribuer un poids d'appariement à chacune de ces paires d'après les règles d'appariement spécifiées et d'utiliser les courbes de distribution des poids des paires simulées pour estimer l'erreur.

Mots clés : Appariement de fichiers; taux d'erreurs de couplage; poids d'appariement; seuil de sélection; dossiers médicaux.

1. Introduction

La démarche fondamentale de couplage d'enregistrements établie par Newcombe, Kennedy, Axford et James (1959) et par Fellegi et Sunter (1969) repose sur l'utilisation d'un poids d'appariement pour évaluer la probabilité qu'un appariement soit correct et sur une règle de décision pour classer les paires d'enregistrements. La règle de décision optimale repose sur deux seuils de poids d'appariement pour la sélection (un seuil supérieur au-dessus duquel un couplage est traité comme un vrai appariement et un seuil inférieur sous lequel un couplage est traité comme un non-appariement). Le choix de ces seuils dépend du taux d'erreurs de couplage acceptable préétabli et de la nécessité de réduire au minimum le nombre de couplages de situation indéterminée entre les deux seuils. De nos jours, les praticiens du couplage informatisé utilisent souvent un seuil de sélection unique pour éviter l'intervention manuelle que requiert le traitement des couplages indéterminés. Habituellement, le système prend automatiquement les décisions concernant les couplages après qu'on l'ait « réglé » de façon à respecter le niveau d'erreurs préétabli. Le défi tient au fait que les méthodes courantes de détermination du seuil de sélection et d'estimation des erreurs de couplage peuvent produire des résultats divergents, selon le type d'erreur de couplage, le choix de l'espace de comparaison et la méthode d'estimation.

Le but du présent article est de partager nos connaissances avec les praticiens qui ont besoin d'une méthode pour orienter le choix des couplages et pour estimer l'erreur. Notre étude de cas porte sur des fichiers d'événements médicaux provenant de la Medical Expenditure Panel Survey (MEPS) réalisée aux États-Unis. La MEPS est conçue pour recueillir des données sur les frais médicaux auprès de répondants sélectionnés dans les ménages et auprès des prestataires de soins médicaux. L'objectif est de combiner les données en provenance des deux sources pour produire des estimations annuelles de l'utilisation des services médicaux et des frais médicaux (pour d'autres renseignements sur la MEPS, consulter Agency for Healthcare Research and Quality 2001).

Nous discutons ici du couplage d'enregistrements portant sur trois ensembles de fichiers annuels d'événements médicaux, provenant de la MEPS de 1996, de la MEPS de 1997 et de la MEPS de 1998. Chaque ensemble comprend un fichier des ménages contenant les événements déclarés par les répondants des ménages pour une année particulière et un fichier des prestataires de soins médicaux contenant les données sur les événements correspondants déclarés par les personnes ayant prodigué les soins aux répondants des ménages. Chaque année, environ 50 000 événements médicaux ont été déclarés pour près de 10 000 personnes et environ 15 000 unités personne-prestataire de soins, en moyenne.

Dans l'article de Dizio, Guarnera et Luzi, des modèles de mélanges finis sont utilisés pour détecter les erreurs dues à l'utilisation d'une unité incorrecte de mesure au stade de la collecte des données d'enquête. Dans un contexte multivarié et en supposant que les données suivent une loi normale multivariée, la méthode permet de déterminer quelles variables sont erronées pour une unité échantillonnée particulière. Les auteurs fournissent aussi des diagnostics pour établir l'ordre de priorité des cas qui doivent faire l'objet d'un examen manuel plus approfondi. La méthode proposée est illustrée au moyen d'un exemple portant sur des données simulées et d'un exemple portant sur des données réelles.

Chiu, Yucel, Zanutto et Zaslavsky présentent une méthode d'imputation multiple de variables contextuelles manquantes en vue de leur utilisation en analyse de régression. Pour chaque enrégistrement dans lequel manque la variable, et pour un échantillon d'enregistrements complets, ils sélectionnent des cas appariés d'après un ensemble de variables d'appariement. L'échantillon d'enregistrements complets est alors utilisé pour estimer à un ajustement de la régression pour d'autres variables non incluses dans les variables d'appariement. Les variables contextuelles pour les enrégistrement incomplets font ensuite l'objet d'une imputation multiple. Enfin, les auteurs décrivent une application à l'étude du cancer du côlon et du rectum et utilisent des simulations pour comparer leur approche à trois autres méthodes d'ajustement pour la non-réponse.

Nandram et Choi examinent l'important problème de la non-réponse non-ignorable lors de l'estimation d'une variable de l'état de santé pour petits domaines. Face à une situation où les estimateurs habituels sont biaisés parce que le nombre de non-répondants est trop élevé, ils essaient de tenir compte des différences par modélisation. Nandram et Choi utilisent deux modèles hiérarchiques bayésiens de la non-réponse non-ignorable, un modèle de sélection et un modèle de mélange de schémas d'observation, pour analyser les données sur la santé. Un élément important dans leur modélisation est l'intégration de l'opinion des médecins en ce qui concerne le comportement de non-réponse et la variable des résultats. Les résultats donnent un ajustement exact pour la non-réponse et une meilleure mesure de précision.

Part et Fuller proposent une méthode en vue de réduire la probabilité d'obtenir des poids d'estimation négatifs lorsqu'on utilise un estimateur par la régression. Leur méthode consiste à approximer d'abord les probabilités d'inclusion, sachant les estimations d'Horvitz-Thompson pour un vecteur de variables auxiliaires, puis à utiliser les probabilités d'inclusion approximatifs conditionnelles comme poids initiaux dans un estimateur par la régression. Ils montrent que leur méthode donne de bons résultats dans une étude en simulation. Ils comparent aussi les poids obtenus par leur méthode à ceux obtenus par la programmation quadratique, le raking ratio, une procédure logit et la méthode du maximum de vraisemblance.

La première des trois communications brèves publiées dans le présent numéro, rédigée par Andersson et Thorburn, montre que l'estimateur par la régression optimal peut être exprimé sous forme d'un estimateur par calage avec une fonction de distance choisie convenablement. L'estimateur optimal résultant est asymptotiquement plus efficace que l'estimateur par la régression généralisée (GREG) habituel. Une petite étude en simulation illustre plusieurs situations où l'estimateur optimal est significativement plus efficace que l'estimateur GREG.

Lynn et Gabler étendent les résultats de Gabler, Hader et Lahiri (volume 25, 1999) à l'expression de l'effet de plan dû à la mise en grappes de Kish. Ils donnent une méthode pratique d'estimation de la quantité de Kish à l'étape du plan d'échantillonnage lorsque seul les nombres totaux d'observations et de grappes sont nécessaires.

Meza et Lahiri examinent les limites d'un critère de sélection standard du modèle de régression, c'est-à-dire la statistique de Mallows, quand on l'applique aux modèles de régression à erreur emboîtée. Ils montrent que, si une application directe de la statistique de Mallows peut produire des méthodes de sélection de modèle inefficaces, une transformation appropriée des données pourrait résoudre le problème.

Finalement, nous voudrions vous informer que Harold Mantel occupera dorénavant le nouveau poste de Rédacteur en chef délégué. Harold fait partie du Comité éditorial depuis 15 ans. Son dévouement à la revue a été notable et sa contribution continue au processus éditorial a été de première importance pour assurer le maintien de la haute qualité de *Techniques d'enquête*.

Dans ce numéro

Ce numéro de *Techniques d'enquête* est dédié à Gordon J. Brackstone qui a récemment pris sa retraite de Statistique Canada. Il était Statisticien en chef adjoint du Secteur de l'Informatique et la Méthodologie et a été président du Comité de direction de *Technique d'enquête* à partir de 1987. Son support continu à la revue a toujours été empreint de discernement et visiblement motivé par un désir constant de stimuler la poursuite de standards élevés de pratiques méthodologiques. De plus, il a lui-même produit des articles pour la revue. Nous sommes vraiment reconnaissants envers Gordon J. Brackstone.

Le présent numéro contient huit articles ordinaires traitant de divers sujets et trois communications brèves. Comme nous l'avons mentionné dans le dernier numéro de la revue, nous lançons une nouvelle section qui sera réservée à des communications brèves. Elle contiendra des articles courts, habituellement de quatre pages environ. Ces communications brèves pourraient être consacrées à la présentation de nouvelles idées sans les traiter complètement comme dans un article ordinaire, à des rapports brefs sur des travaux empiriques ou à des discussions ou des compléments d'autres articles publiés dans la revue.

Depuis quatre ans, le numéro de juin de *Techniques d'enquête* contient un article invité en l'honneur de Joseph Waksberg. À partir de cette année, cet article invité sera publié dans le numéro de décembre de la revue afin de mieux le synchroniser avec la présentation correspondante qui est faite au symposium annuel de Statistique Canada sur la méthodologie à l'automne. L'auteur Waksberg de cette année est J.N.K. Rao et son article sera sur « l'interaction entre la théorie et les méthodes d'échantillonnage : Une évaluation ».

Dans l'article d'ouverture du présent numéro, Winglee, Valliant et Scheuren présentent une nouvelle approche de simulation pour estimer les taux d'erreurs pour la sélection des seuls lors des couplages d'enregistrements. Pour chaque paire susceptible d'être un appariement vrai, il existe un vecteur de résultats de comparaisons qui détermine le poids d'appariement. Les auteurs supposent que chaque résultat de comparaison correspond à un modèle multinomial, et que la loi multinomiale diffère pour les appariements vrais et les non-appariements. Ils estiment les lois d'après un échantillon, puis les utilisent pour simuler les lois de probabilité des poids appariés pour les appariements vrais et les non-appariements. Ils illustrent la méthode au moyen d'une étude de cas en se servant de données provenant de la Medical Expenditure Panel Survey (MEPS) réalisée aux États-Unis.

Krewski, Dewanjy, Wang, Bartlett, Zielinski et Mallick étudient les effets des erreurs de couplage d'enregistrements, aussi bien les résultats faussement positifs que faussement négatifs, sur les estimations du risque dans les études de cohorte. Ils montrent analytiquement comment les erreurs de couplage introduisent un biais et une variabilité supplémentaire dans les nombres observés et attendus de décès, ainsi que dans les estimations des ratios standardisés de mortalité et des coefficients de régression du risque relatif. Ils discutent des résultats dans leurs conclusions et soulignent les travaux qui devraient être réalisés dans ce domaine.

L'article rédigé par van den Brakel et Renssen traite du problème de la vérification d'hypothèses sous différentes mises en œuvre de l'enquête, comme des conceptions différentes du questionnaire, lorsqu'on utilise un plan d'échantillonnage complexe. Ils élaborent une théorie fondée sur le plan de sondage pour les cas où les diverses mises en œuvre de l'enquête sont affectées à des sous-échantillons au moyen de plans d'expérience en randomisation totale ou en blocs randomisés. La théorie s'appuie aussi sur des modèles de l'erreur de mesure. Les auteurs utilisent la statistique de Wald fondée sur le plan de sondage pour comparer les diverses mises en œuvre de l'enquête.

Tsuchiya aborde d'une manière intéressante l'ancien problème que soulevaient les questions délicates posées lors des enquêtes. Au lieu d'utiliser la technique de réponse aléatoire qui offre peu de contrôle au chercheur, il propose d'adapter la technique du dénombrement d'items au cas des questions délicates. La technique de dénombrement d'items consiste à présenter à l'enquêteur une liste de plusieurs phrases et de lui demander de choisir toutes celles qui s'appliquent à lui. Le chercheur construit la liste de deux façons : la première contient la phrase délicate, tandis que la deuxième ne la contient pas. Tsuchiya présente divers estimateurs pour cette technique et donne un exemple intéressant ayant trait au caractère national japonais.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 31, numéro 1, juin 2005

Table des matières

Dans ce numéro	1
M. Winglee, R. Valliant et F. Scheuren Une étude de cas en couplage d'enregistrements	3
D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J. M. Zielinski et R. Mallick L'effet des erreurs de couplage d'enregistrements sur les estimations du risque dans les études-cohorte de mortalité	15
Jan A. van den Brakel et Robert H. Renssen Analyse d'expériences intégrées dans des plans de sondage complexes	25
Takahiro Tsuchiya Estimateurs de domaine pour la technique du dénombrement d'items	45
Marco Di Zio, Ugo Guamera et Orfetta Luzi Vérification des erreurs systématiques d'unité de mesure au moyen de la modélisation par mélanges	57
Wai Fung Chiu, Recai M. Yucel, Elaine Zanutto et Alan M. Zaslavsky Utilisation de substituts appariés pour améliorer les imputations dans les bases de données couplées géographiquement	69
Balagobin Nandram et Jai Won Choi Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : Une application aux données de la NHANES	79
Mingue Park et Wayne A. Fuller Vers des poids de régression non négatifs pour les échantillons d'enquête	93
Communications brèves	
Per Gösta Andersson et Daniel Thorburn Une distance de calage optimale menant à un estimateur par la régression optimal	103
Peter Lynn et Siegfried Gabler Approximations de b^* dans la prévision des effets du plan dus à la mise en grappes	109
Jane L. Meza et P. Lahiri Une note sur la statistique C_p sous un modèle de régression à erreur emboîtée	115

A Gordon J. Brackstone

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président D. Royce

Anciens présidents G.J. Brackstone
R. Platek

Membres J. Gambino

J. Kovar
H. Mantel
M.P. Singh

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, *Statistique Canada*

Rédacteur en chef délégué H. Mantel, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*

D.A. Binder, *Statistique Canada*

J.M. Brick, *Westat, Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J.L. Eitinger, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Low State University*

J. Gambino, *Statistique Canada*

M.A. Hiddigton, *Office for National Statistics*

G. Kalton, *Westat, Inc.*

P. Kott, *National Agricultural Statistics Service*

J. Kovar, *Statistique Canada*

P. Lahiri, *JPSM, University of Maryland*

G. Nathan, *Hebrew University*

D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*

Rédacteurs adjoints J.-F. Beaumont, P. Dick et W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception de contrats d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et l'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, Dr. M.P. Singh, singhmp@statcan.ca (Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Pré Tunney, Ottawa, Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ x 2 exemplaires), autres pays, 30 \$ CA (15 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commander par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Juillet 2005

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrément sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2005

Publication autorisée par le ministre
responsable de Statistique Canada

JUN 2005 • VOLUME 31 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

•

VOLUME 31

•

JUIN 2005

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE



12-001



Government
Publications

SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2005

•

VOLUME 31

•

NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2005 • VOLUME 31 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2006

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

February 2006

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Members J. Gambino
J. Kovar
H. Mantel

Past Chairmen G.J. Brackstone
R. Platek

E. Rancourt (Production Manager)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Deputy Editor H. Mantel, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*

D.A. Binder, *Statistics Canada*

J.M. Brick, *Westat, Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J.L. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

M.A. Hidirolou, *Office for National Statistics*

G. Kalton, *Westat, Inc.*

P. Kott, *National Agricultural Statistics Service*

J. Kovar, *Statistics Canada*

P. Lahiri, *JPSM, University of Maryland*

G. Nathan, *Hebrew University*

D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

M. Thompson, *University of Waterloo*

Y. Tillé, *Université de Neuchâtel*

R. Valliant, *JPSM, University of Michigan*

V.J. Verma, *Università degli Studi di Siena*

J. Waksberg, *Westat, Inc.*

K.M. Wolter, *Iowa State University*

A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (rte@statcan.ca, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

Survey Methodology
A journal Published by Statistics Canada
Volume 31, Number 2, December 2005

Contents

In This Issue	111
In Memoriam M.P. Singh	113
Waksberg Invited Paper Series	
J.N.K. Rao Interplay Between Sample Survey Theory and Practice: An Appraisal	117
Regular Papers	
Wayne A. Fuller and Jae Kwang Kim Hot Deck Imputation for the Response Model	139
J. Michael Brick, Michael E. Jones, Graham Kalton and Richard Valliant Variance Estimation with Hot Deck Imputation: A Simulation Study of Three Methods	151
Roderick J. Little and Sonya Vartivarian Does Weighting for Nonresponse Increase the Variance of Survey Means?	161
Alistair James O'Malley and Alan Mark Zaslavsky Variance-Covariance Functions for Domain Means of Ordinal Survey Items	169
Bharat Bhushan Singh, Girja Kant Shukla and Debasis Kundu Spatio-Temporal Models in Small Area Estimation	183
Liv Belsby, Jan Bjørnstad and Li-Chun Zhang Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey	197
Balgobin Nandram, Lawrence H. Cox and Jai Won Choi Bayesian Analysis of Nonignorable Missing Categorical Data: An Application to Bone Mineral Density and Family Income	213
Short Notes	
Jean-François Beaumont On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment	227
Alfredo Bustos On the Correlation Structure of Sample Units	233
Changbao Wu Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling	239
Acknowledgements	245

In This Issue

It is with great sadness that we note the recent passing of M.P. Singh, Editor of the *Survey Methodology* journal since the very first issue in 1975. This issue of the journal opens with a brief obituary in memoriam.

This issue of *Survey Methodology* also contains the fifth paper in the annual invited paper series in honour of Joseph Waksberg. A short biography of Joseph Waksberg was given in the June 2001 issue of the journal, along with the first paper in the series. I would like to thank the members of the selection committee- Michael Brick, chair, David Bellhouse, Gordon Brackstone and Paul Biemer – for having selected Jon Rao as the author of this year's Waksberg paper.

In his paper entitled "Interplay Between Sample Survey Theory and Practice: An Appraisal", Rao traces how survey methods are stimulated by new theoretical developments, and how theory is challenged by survey practice. After summarizing fifty years of contributions from 1920 to 1970, he presents more detailed discussions of more recent developments in several areas. Finally, he discusses several examples of important theory that is not yet widely applied in practice.

In their paper, Fuller and Kim develop and study an efficient hot-deck imputation method under the assumption that response probabilities are equal within imputation cells. Their proposed method is based on the idea of fractional imputation and uses regression techniques to obtain an approximation of the fully efficient version of fractional imputation. Variance estimation is developed for replication methods. Their proposed method is shown to work well in a simulation study.

The paper by Brick, Jones, Kalton and Valliant compares through a simulation study three variance estimation methods in the presence of hot-deck imputation: the model-assisted method, the adjusted jackknife method and multiple imputation. The goal of the simulation study is to study the properties of these variance estimators when their underlying assumptions do not hold. They found that the coverage rate of confidence intervals is not close to the nominal level when the point estimates are biased due failure to take into account the domains of interest at the imputation stage. They conclude by noting that the differences between the variance estimators were too small and inconsistent to support claims that any one of them is superior in general.

Little and Vartivarian study the effect of nonresponse weighting on the Mean Squared Error (MSE) of a population mean estimator. Nonresponse weighting adjustments are obtained by adjusting design weights by the inverse of response rates within cells. They come to the conclusion that a covariate must have two characteristics to reduce nonresponse bias: it needs to be related to both the probability of response and to the survey outcome. If the latter is true, nonresponse weighting can also reduce nonresponse variance. Estimates of the MSE are proposed and used to define a composite estimator. This composite estimator worked well when evaluated in a simulation study.

O'Malley and Zaslavsky present generalized variance-covariance modeling functions (GVCFs) for multivariate means of ordinal survey items, for both complete data and data with structured non-response. After developing and evaluating their methods, they give an illustration using data from the Consumer Assessments of Health Plans Study. In the concluding section they discuss some issues related to the application of GVCFs.

The paper by Singh, Shukla and Kundu develops spatial and spatial-temporal models for small area estimation, as well as estimation of the MSE of the resulting EBLUPs. The models are applied to monthly per capita consumption expenditure data, and they conclude that the models can be very effective when there are significant correlations due to neighborhood effects.

Belsby, Bjørnstad and Zhang discuss modeling to estimate the number of households of different sizes when there is nonignorable nonresponse. They model the response mechanism conditional on household size, using registered family size as supplementary data. After developing their modeling approach, they produce and evaluate estimates using data from the 1992 Norwegian Consumer Expenditure Survey.

Nandram, Cox and Choi consider an analysis for categorical data from a single two-way table with both item and unit nonresponse or, in their terminology, partial classification. They propose to use a Bayesian approach for modeling different patterns of missingness under ignorability and non-ignorability assumptions. The methods are illustrated using incompletely-observed bivariate data from the National Health and Nutrition Examination Survey where the variables subject to missingness are bone mineral density and family income.

In the first of three short notes in this issue, Beaumont discusses the use of data collection process information in nonresponse weight adjustment. He then presents an example from the Canadian Labour Force Survey using the number of attempts to contact a survey unit. An important result is that if the collection process information can be treated as random, then this approach does not introduce any bias.

Starting from basic principles, Bustos derives an explicit form for the probability function of an ordered sample. Using this function, he shows how it can be used to compute inclusion probabilities with illustrations for common sample designs. Finally, he gives the general form for the correlation matrix of sample units, which depends solely on the inclusion probabilities.

Finally, the paper by Wu briefly reviews some theory about the Pseudo Empirical Likelihood (PEL) method in survey sampling, and presents algorithms for computing maximum PEL estimators and for constructing PEL ratio confidence intervals. Functions using the statistical software R and S-PLUS are given to help implement these algorithms in real surveys or in simulation studies.

Harold Mantel

In Memoriam M.P. Singh (1941-2005)

Dr. Mangala P. Singh was born in India on December 26th, 1941 and received his PhD in 1969 from the Indian Statistical Institute, with a specialization in survey sampling. He joined Statistics Canada in 1970, where he rose to the position of Director of Household Survey Methods Division in 1994, a position he held at his death on August 24th, 2005.

M.P., as he was known to everyone, was a leading figure in the application of statistical methods at Statistics Canada. He was probably most closely associated with the Labour Force Survey, one of the agency's most important surveys. He directed the methodology of the LFS through redesigns in the 1970s, 1980s, 1990s and early 21st century, introducing innovations at every turn, but always ensuring that changes were well-tested and sound. In the later years of his career, he also oversaw the development of several new and innovative health surveys and directed the development of statistical programs in the areas of household expenditures, education and justice.

M.P.'s role as the Editor-in-Chief of the journal *Survey Methodology* had a transformative effect on the profession of survey methodology, both in Canada and abroad. M.P. was the founding editor of the journal, and for 30 years he guided its evolution into a flagship publication of Statistics Canada. Thanks to his ability to attract a stellar team of associate editors and contributors, *Survey Methodology* is now recognized as one of the pre-eminent journals of its kind in the world. Even in recent years, M.P. continued to introduce innovations such as the Waksberg series of papers and electronic publishing.

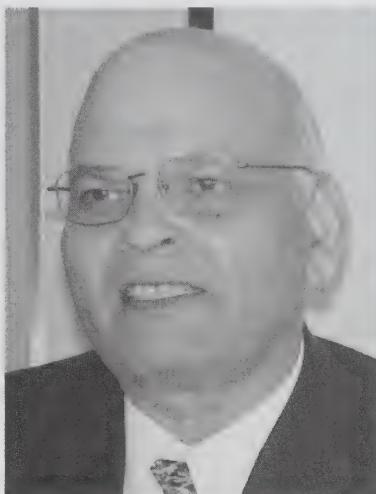
M.P. was a source of many other "big ideas" throughout his career at Statistics Canada. During the 1970s he was instrumental in gaining support for the idea of stable funding for methodology research, and he personally chaired the Methodology Research and Development

Committee in its formative years. He encouraged numerous researchers and went out of his way to make them feel at home at Statistics Canada. Turning 60 did not stem the flow of ideas in any way. M.P. devoted considerable energy in the past four years to his proposal for a major overhaul of the way household surveys are conducted in Canada. As a result of his efforts, people throughout Statistics Canada are working on ways to implement his vision, and his influence on Canada's household surveys will be felt for many years.

M.P. had a special love for statistical research and for statistics as a profession. He personally authored over 40 papers in international journals, co-edited two books published by Wiley and Sons, and organized sessions and presented papers at numerous statistical conferences. He served on various committees and task forces of the Statistical Society of Canada, the International Statistical Institute and the American Statistical Association. He also served as Secretary of Statistics Canada's external Advisory Committee on Statistical Methods. In turn, the profession honoured him; he was elected to the International Statistical Institute in 1975, and in 1988 he became a Fellow of the American Statistical Association.

However it is his influence on an entire generation of statisticians that may be his greatest legacy. He was a mentor, a coach, a patriarch and a friend to all who knew him. He inspired others to give their best, and they did. He was always ready with a laugh, a smile and a friendly word of encouragement. He dedicated his life to the profession of statistics and it is through those whom he touched that his true contribution is measured.

He is survived by his wife Savitri, his two daughters Mala and Mamta, and his son Rahul.



Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work. The author receives a cash award made possible by a grant from Westat, in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially by the American Statistical Association. Previous winners were Gad Nathan, Wayne Fuller, Tim Holt, Norman Bradburn, Jon Rao, and Alastair Scott. The first five papers in the series have already appeared in *Survey Methodology*.

Previous Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)

Nominations:

The author of the 2007 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, Gordon Brackstone, 78 Charing Road, Ottawa, Ontario, Canada, K2G 4C9, by email to Gordon.brackstone@sympatico.ca or by fax 1-613-951-1394. Nominations and suggestions for topics must be received by February 28, 2006.

2005 Waksberg Invited Paper

Author: J.N.K. Rao

J.N.K. Rao is Distinguished Research Professor at Carleton University, Ottawa. He has published many articles on a wide range of topics in survey sampling theory and methods and he is the author of the 2003 Wiley book "Small Area Estimation". His research interests in survey sampling include analysis of survey data, small area estimation, missing data and imputation, re-sampling methods and empirical likelihood inference. His 1981 JASA paper (with A.J. Scott) on analysis of survey data was selected as a landmark paper in survey sampling theory and methods. He has been a Member of the Advisory Committee on Statistical Methods of Statistics Canada since its inception 20 years ago. He is a Fellow of the Royal Society of Canada and received the 1994 Gold Medal of the Statistical Society of Canada.

Members of the Waskberg Paper Selection Committee (2005-2006)

Gordon Brackstone, (Chair)

Wayne Fuller, *Iowa State University*

Sharon Lohr, *Arizona State University*

Past Chairs:

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Interplay Between Sample Survey Theory and Practice: An Appraisal

J.N.K. Rao¹

Abstract

A large part of sample survey theory has been directly motivated by practical problems encountered in the design and analysis of sample surveys. On the other hand, sample survey theory has influenced practice, often leading to significant improvements. This paper will examine this interplay over the past 60 years or so. Examples where new theory is needed or where theory exists but is not used will also be presented.

Key Words: Analysis of survey data; Early contributions; Inferential issues; Re-sampling methods; Small area estimation.

1. Introduction

In this paper I will examine the interplay between sample survey theory and practice over the past 60 years or so. I will cover a wide range of topics: early landmark contributions that have greatly influenced practice, inferential issues, calibration estimation that ensures consistency with user specified totals of auxiliary variables, unequal probability sampling without replacement, analysis of survey data, the role of resampling methods, and small area estimation. I will also present some examples where new theory is needed or where theory exists but is not used widely.

2. Some Early Landmark Contributions: 1920 – 1970

This section gives an account of some early landmark contributions to sample survey theory and methods that have greatly influenced the practice. The Norwegian statistician A.N. Kiaer (1897) is perhaps the first to promote sampling (or what was then called “the representative method”) over complete enumeration, although the oldest reference to sampling can be traced back to the great Indian epic Mahabharata (Hacking 1975, page 7). In the representative method the sample should mirror the parent finite population and this may be achieved either by balanced sampling through purposive selection or by random sampling. The representative method was used in Russia as early as 1900 (Zarkovic 1956) and Wright conducted sample surveys in the United States around the same period using this method. By the 1920s, the representative method was widely used, and the International Statistical Institute played a prominent role by creating a committee in 1924 to report on the representative method. This committee’s report discussed theoretical and practical aspects of the random sampling method. Bowley’s (1926) contribution to this report includes his fundamental work on stratified random

sampling with proportional allocation, leading to a representative sample with equal inclusion probabilities. Hubback (1927) recognized the need for random sampling in crop surveys: “The only way in which a satisfactory estimate can be found is by as close an approximation to random sampling as the circumstances permit, since that not only gets rid of the personal limitations of the experimenter but also makes it possible to say what is the probability with which the results of a given number of samples will be within a given range from the mean. To put this into definite language, it should be possible to find out how many samples will be required to secure that the odds are at least 20:1 on the mean of the samples within one maund of the true mean”. This statement contains two important observations on random sampling: (1). It avoids personal biases in sample selection. (2). Sample size can be determined to satisfy a specified margin of error apart from a chance of 1 in 20. Mahalanobis (1946b) remarked that R.A. Fisher’s fundamental work at Rothamsted Experimental Station on design of experiments was influenced directly by Hubback (1927).

Neyman’s (1934) classic landmark paper laid the theoretical foundations to the probability sampling (or design-based) approach to inference from survey samples. He showed, both theoretically and with practical examples, that stratified random sampling is preferable to balanced sampling because the latter can perform poorly if the underlying model assumptions are violated. Neyman also introduced the ideas of efficiency and optimal allocation in his theory of stratified random sampling without replacement by relaxing the condition of equal inclusion probabilities. By generalizing the Markov theorem on least squares estimation, Neyman proved that the stratified mean, $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, is the best estimator of the population mean, $\bar{Y} = \sum_h W_h \bar{Y}_h$, in the linear class of unbiased estimators of the form $\bar{y}_b = \sum_h W_h \sum_i b_{hi} y_{hi}$, where W_h , \bar{y}_h and \bar{Y}_h are

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

the h^{th} stratum weight, sample mean and population mean ($h=1, \dots, L$), and b_{hi} is a constant associated with the item value y'_{hi} observed on the i^{th} sample draw ($i=1, \dots, n_h$) in the h^{th} stratum. Optimal allocation (n_1, \dots, n_L) of the total sample size, n , was obtained by minimizing the variance of \bar{y}_{st} subject to $\sum_h n_h = n$; an earlier proof of Neyman allocation by Tschuprow (1923) was later discovered. Neyman also proposed inference from larger samples based on normal theory confidence intervals such that the frequency of errors in the confidence statements based on all possible stratified random samples that could be drawn does not exceed the limit prescribed in advance "*whatever the unknown properties of the population*". Any method of sampling that satisfies the above frequency statement was called "representative". Note that Hubback (1927) earlier alluded to the frequency statement associated with the confidence interval. Neyman's final contribution to the theory of sample surveys (Neyman 1938) studied two-phase sampling for stratification and derived the optimal first phase and second phase sample sizes, n' and n , by minimizing the variance of the estimator subject to a given cost $C = n'c' + nc$, where the second phase cost per unit, c , is large relative to the first phase cost per unit, c' .

The 1930's saw a rapid growth in demand for information, and the advantages of probability sampling in terms of greater scope, reduced cost, greater speed and model-free features were soon recognized, leading to an increase in the number and type of surveys taken by probability sampling and covering large populations. Neyman's approach was almost universally accepted by practicing survey statisticians. Moreover, it inspired various important extensions, mostly motivated by practical and efficiency considerations. Cochran's (1939) landmark paper contains several important results: the use of ANOVA to estimate the gain in efficiency due to stratification, estimation of variance components in two-stage sampling for future studies on similar material, choice of sampling unit, regression estimation under two-phase sampling and effect of errors in strata sizes. This paper also introduced the super-population concept: "The finite population should itself be regarded as a random sample from some infinite population". It is interesting to note that Cochran at that time was critical of the traditional fixed population concept: "Further, it is far removed from reality to regard the population as a fixed batch of known numbers". Cochran (1940) introduced ratio estimation for sample surveys, although an early use of the ratio estimator dates back to Laplace (1820). In another landmark paper (Cochran 1942), he developed the theory of regression estimation. He derived the conditional variance of the usual regression estimator for a fixed sample and also a sample estimator of this variance, assuming a linear regression model $y = \alpha + \beta x + e$, where e has mean zero and

constant variance in arrays in which x is fixed. He also noted that the regression estimator remains (model) unbiased under non-random sampling, provided the assumed linear regression model is correct. He derived the average bias under model deviations (in particular, quadratic regression) for simple random sampling as the sample size n increased. Cochran then extended his results to weighted regression and derived the now well-known optimality result for the ratio estimator, namely it is a "best unbiased linear estimate if the mean value and variance both change proportional to x ". The latter model is called the ratio model in the current literature. Madow and Madow (1944) and Cochran (1946) compared the expected (or anticipated) variance under a super-population model to study the relative efficiency of systematic sampling and stratified random sampling analytically. This paper stimulated much subsequent research on the use of super-population models in the choice of probability sampling strategies, and also for model-dependent and model-assisted inferences (see section 3).

In India, Mahalanobis made pioneering contributions to sampling by formulating cost and variance functions for the design of surveys. His 1944 landmark paper (Mahalanobis 1944) provides deep theoretical results on the efficient design of sample surveys and their practical applications, in particular to crop acreage and yield surveys. The well-known optimal allocation in stratified random sampling with cost per unit varying across strata is obtained as a special case of his general theory. As early as 1937, Mahalanobis used multi-stage designs for crop yield surveys with villages, grids within villages, plots within grids and cuts of different sizes and shapes as sampling units in the four stages of sampling (Murthy 1964). He also used a two-phase sampling design for estimating the yield of cinchona bark. He was instrumental in establishing the National Sample Survey (NSS) of India, the largest multi-subject continuing survey operation with full-time staff using personal interviews for socioeconomic surveys and physical measurements for crop surveys. Several prominent survey statisticians, including D.B. Lahiri and M.N. Murthy, were associated with the NSS.

P.V. Sukhatme, who studied under Neyman, also made pioneering contributions to the design and analysis of large-scale agricultural surveys in India, using stratified multi-stage sampling. Starting in 1942–1943 he developed efficient designs for the conduct of nationwide surveys on wheat and rice crops and demonstrated high degree of precision for state estimates and reasonable margin of error for district estimates. Sukhatme's approach differed from that of Mahalanobis who used very small plots for crop cutting employing *ad hoc* staff of investigators. Sukhatme (1947) and Sukhatme and Panse (1951) demonstrated that

the use of a small plot might give biased estimates due to the tendency of placing boundary plants inside the plot when there is doubt. They also pointed out that the use of an *ad hoc* staff of investigators, moving rapidly from place to place, forces the plot measurements on only those sample fields that are ready for harvest on the date of the visit, thus violating the principle of random sampling. Sukhatme's solution was to use large plots to avoid boundary bias and to entrust crop-cutting work to the local revenue or agricultural agency in a State.

Survey statisticians at the U.S. Census Bureau, under the leadership of Morris Hansen, William Hurwitz, William Madow and Joseph Waksberg, made fundamental contributions to sample survey theory and practice during the period 1940–70, and many of those methods are still widely used in practice. Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn with probability proportional to size measure (PPS sampling) and then sub-sampled at a rate that ensures self-weighting (equal overall probabilities of selection) within strata. This approach provides approximately equal interviewer work loads which is desirable in terms of field operations. It also leads to significant variance reduction by controlling the variability arising from unequal PSU sizes without actually stratifying by size and thus allowing stratification on other variables to reduce the variance. On the other hand, workloads can vary widely if the PSUs are selected by simple random sampling and then sub-sampled at the same rate within each stratum. PPS sampling of PSUs is now widely used in the design of large-scale surveys, but two or more PSUs are selected without replacement from each stratum such that the PSU inclusion probabilities are proportional to size measures (see section 5).

Many large-scale surveys are repeated over time, such as the monthly Canadian Labour Force Survey (LFS) and the U.S. Current Population Survey (CPS), with partial replacement of ultimate units (also called rotation sampling). For example, in the LFS the sample of households is divided into six rotation groups (panels) and a rotation group remains in the sample for six consecutive months and then drops out of the sample, thus giving five-sixth overlap between two consecutive months. Yates (1949) and Patterson (1950), following the initial work of Jessen (1942) for sampling on two occasions with partial replacement of units, provided the theoretical foundations for design and estimation of repeated surveys, and demonstrated the efficiency gains for level and change estimation by taking advantage of past data. Hansen, Hurwitz, Nisselson and Steinberg (1955) developed simpler estimators, called K – composite estimators, in the context of stratified multi-stage designs with PPS sampling in the first stage. Rao and

Graham (1964) studied optimal replacement policies for the K – composite estimators. Various extensions have also been proposed. Composite estimators have been used in the CPS and other continuing large scale surveys. Only recently, the Canadian LFS adopted a type of composite estimation, called regression composite estimation, that makes use of sample information from previous months and that can be implemented with a regression weights program (see section 4).

Keyfitz (1951) proposed an ingenious method of switching to better PSU size measures in continuing surveys based on the latest census counts. His method ensures that the probability of overlap with the previous sample of one PSU per stratum is maximized, thus reducing the field costs and at the same time achieving increased efficiency by using the better size measures in PPS sampling. The Canadian LFS and other continuing surveys have used the Keyfitz method. Raj (1956) formulated the optimization problem as a “transportation problem” in linear programming. Kish and Scott (1971) extended the Keyfitz method to changing strata and size measures. Ernst (1999) has given a nice account of the developments over the past 50 years in sample co-ordination (maximizing or minimizing the sample overlap) using transportation algorithms and related methods; see also Mach, Reiss and Schiopu-Kratina (2005) for applications to business surveys with births and deaths of firms.

Dalenius (1957, Chapter 7) studied the problem of optimal stratification for a given number of strata, L , under the Neyman allocation. Dalenius and Hodges (1959) obtained a simple approximation to optimal stratification, called the \sqrt{f} rule, which is extensively used in practice. For highly skewed populations with a small number of units accounting for a large share of the total Y , such as business populations, efficient stratification requires one take-all stratum ($n_1 = N_1$) of big units and take-some strata of medium and small size units. Lavallée and Hidioglou (1988) and Rivest (2002) developed algorithms for determining the strata boundaries using power allocation (Fellegi 1981; Bankier 1988) and Neyman allocation for the take some strata. Statistics Canada and other agencies currently use those algorithms for business surveys.

The focus of research prior to 1950 was on estimating population totals and means for the whole population and large planned sub-populations, such as states or provinces. However, users are also interested in totals and means for unplanned sub-populations (also called domains) such as age-sex groups within a province, and parameters other than totals and means such as the median and other quantiles, for example median income. Hartley (1959) developed a simple, unified theory for domain estimation applicable to any design, requiring only the standard formulae for the estimator of total and its variance estimator, denoted in the

operator notation as $\hat{Y}(y)$ and $v(y)$ respectively. He introduced two synthetic variables ${}_j y_i$ and ${}_j a_i$ which take the values y_j and 1 respectively if the unit i belongs to domain j and equal to 0 otherwise. The estimators of domain total ${}_j Y = Y({}_j y)$ and domain size ${}_j N = Y({}_j a)$ are then simply obtained from the formulae for $\hat{Y}(y)$ and $v(y)$ by replacing y_i by ${}_j y_i$ and ${}_j a_i$ respectively. Similarly, estimators of domain means and domain differences and their variance estimators are obtained from the basic formulae for $\hat{Y}(y)$ and $v(y)$. Durbin (1968) also obtained similar results. Domain estimation is now routinely done using Hartley's ingenious method.

For inference on quantiles, Woodruff (1952) proposed a simple and ingenious method of getting a $(1-\alpha)$ -level confidence interval under general sampling designs, using only the estimated distribution function and its standard error (see Lohr's (1999) book, pages 311–313). Note that the latter are simply obtained from the formulae for a total by changing y to an indicator variable. By equating the Woodruff interval to a normal theory interval on the quantile, a simple formula for the standard error of the p^{th} quantile estimator may also be obtained as half the length of the interval divided by the upper $\alpha/2$ -point of the standard $N(0, 1)$ distribution which equals 1.96 if $\alpha = 0.05$ (Rao and Wu 1987; Francisco and Fuller 1991). A surprising property of the Woodruff interval is that it performs well even when p is small or large and sample size is moderate (Sitter and Wu 2001).

The importance of measurement errors was realized as early as the 1940s. Mahalanobis' (1946a) influential paper developed the technique of interpenetrating sub-samples (called replicated sampling by Deming 1960). This method was extensively used in large-scale sample surveys in India for assessing both sampling and measurement errors. The sample is drawn in the form of two or more independent sub-samples according to the same sampling design such that each sub-sample provides a valid estimate of the total or mean. The sub-samples are assigned to different interviewers (or teams) which leads to a valid estimate of the total variance that takes proper account of the correlated response variance component due to interviewers. Interpenetrating sub-samples increase the travel costs of interviewers, but they can be reduced through modifications of interviewer assignments. Hansen, Hurwitz, Marks and Mauldin (1951), Sukhatme and Seth (1952) and Hansen, Hurwitz and Bershad (1961) developed basic theories under additive measurement error models, and decomposed the total variance into sampling variance, simple response variance and correlated response variance. The correlated response variance due to interviewers was shown to be of the order k^{-1} regardless of the sample size, where k is the number of interviewers. As a result, it can dominate the total variance

if k is not large. The 1950 U.S. Census interviewer variance study showed that this component was indeed large for small areas. Partly for this reason, self-enumeration by mail was first introduced in the 1960 U.S. Census to reduce this component of the variance (Waksberg 1998). This is indeed a success story of theory influencing practice. Fellegi (1964) proposed a combination of interpenetration and replication to estimate the covariance between sampling and response deviations. This component is often neglected in the decomposition of total variance but it could be sizeable in practice.

Yet another early milestone in sample survey methods is the concept of design effect (DEFF) due to Leslie Kish (see Kish 1965, section 8.2). The design effect is defined as the ratio of the actual variance of a statistic under the specified design to the variance that would be obtained under simple random sampling of the same size. This concept is especially useful in the presentation and modeling of sampling errors, and also in the analysis of complex survey data involving clustering and unequal probabilities of selection (see section 6).

We refer the reader to Kish (1995), Kruskal and Mosteller (1980), Hansen, Dalenius and Tepping (1985) and O'Muircheartaigh and Wong (1981) for reviews of early contributions to sample survey theory and methods.

3. Inferential Issues

3.1 Unified Design-Based Framework

The development of early sampling theory progressed more or less inductively, although Neyman (1934) studied best linear unbiased estimation for stratified random sampling. Strategies (design and estimation) that appeared reasonable were entertained and relative properties were carefully studied by analytical and/or empirical methods, mainly through comparisons of mean squared errors, and sometimes also by comparing anticipated mean squared errors or variances under plausible super-population models, as noted in section 2. Unbiased estimation under a given design was not insisted upon because it "often results in much larger mean squared error than necessary" (Hansen, Hurwitz and Tepping 1983). Instead, design consistency was deemed necessary for large samples *i.e.*, the estimator approaches the population value as the sample size increases. Classical text books by Cochran (1953), Deming (1950), Hansen, Hurwitz and Madow (1953), Sukhatme (1954) and Yates (1949), based on the above approach, greatly influenced survey practice. Yet, academic statisticians paid little attention to traditional sampling theory, possibly because it lacked a formal theoretical framework and was not integrated with mainstream statistical theory. Numerous prestigious statistics departments in North America did not offer graduate courses in sampling theory.

Formal theoretical frameworks and approaches to integrating sampling theory with mainstream statistical inference were initiated in the 1950s under a somewhat idealistic set-up that focussed on sampling errors assuming the absence of measurement or response errors and non-response. Horvitz and Thompson (1952) made a basic contribution to sampling with arbitrary probabilities of selection by formulating three subclasses of linear design-unbiased estimators of a total Y that include the Markov class studied by Neyman as one of the subclasses. Another subclass with design weight d_i attached to a sample unit i and depending only on i admitted the well-known estimator with weight inversely proportional to the inclusion probability π_i as the only unbiased estimator. Narain (1951) also discovered this estimator, so it should be called the Narain-Horvitz-Thompson (NHT) estimator rather than the HT estimator as it is commonly known. For simple random sampling, the sample mean is the best linear unbiased estimator (BLUE) of the population mean in the three subclasses, but this is not sufficient to claim that the sample mean is the best in the class of all possible linear unbiased estimators. Godambe (1955) proposed a general class of linear unbiased estimators of a total Y by recognizing the sample data as $\{(i, y_i), i \in s\}$ and by letting the weight depend on the sample unit i as well as on the other units in the sample s , that is, the weight is of the form $d_i(s)$. He then established that the BLUE does not exist in the general class

$$\hat{Y} = \sum_{i \in s} d_i(s) y_i, \quad (1)$$

even under simple random sampling. This important negative theoretical result was largely overlooked for about 10 years. Godambe also established a positive result by relating y to a size measure x using a super-population regression model through origin with error variance proportional to x^2 , and then showing that the NHT estimator under any fixed sample size design with π_i proportional to x_i minimizes the anticipated variance in the unbiased class (1). This result clearly shows the conditions on the design for the use of the NHT estimator. Rao (1966) recognized the limitations of the NHT estimator in the context of surveys with PPS sampling and multiple characteristics. Here the NHT estimator will be very inefficient when a characteristic y is unrelated or weakly related to the size measure x (such as poultry count y and farm size x in a farm survey). Rao proposed efficient alternative estimators for such cases that ignore the NHT weights. Ignoring the above results, some theoretical criteria were later advanced in the sampling literature to claim that the NHT estimator should be used for *any* sampling design. Using an amusing example of circus elephants, Basu (1971) illustrated the futility of such criteria. He constructed a "bad" design with π_i unrelated to y_i and then demonstrated that the NHT estimator leads to absurd

estimates which prompted the famous mainstream Bayesian statistician Dennis Lindley to conclude that this counterexample destroys the design-based sample survey theory (Lindley 1996). This is rather unfortunate because NHT and Godambe clearly stated the conditions on the design for a proper use of the NHT estimator, and Rao (1966) and Hajek (1971) proposed alternative estimators to deal with multiple characteristics and bad designs, respectively. It is interesting to note that the same theoretical criteria led to a bad variance estimator of the NHT estimator as the 'optimal' choice (Rao and Singh 1973).

Attempts were also made to integrate sample survey theory with mainstream statistical inference via the likelihood function. Godambe (1966) showed that the likelihood function from the sample data $\{(i, y_i), i \in s\}$, regarding the N - vector of unknown y - values as the parameter, provides no information on the unobserved sample values and hence on the total Y . This uninformative feature of the likelihood function is due to the label property that treats the N population units as essentially N post-strata. A way out of this difficulty is to take the Bayesian route by assuming informative (exchangeable) priors on the parameter vector (Ericson 1969). An alternative route (design-based) is to ignore some aspects of the sample data to make the sample non-unique and thus arrive at an informative likelihood function (Hartley and Rao 1968; Royall 1968). For example, under simple random sampling, suppressing the labels i and regarding the data as $\{(i, y_i), i \in s\}$ in the absence of information relating i to y_i , leads to the sample mean as the maximum likelihood estimator of the population mean. Bayesian estimation, assuming non-informative prior distributions, leads to results similar to Ericson's (1969) but depends on the sampling design unlike Ericson's. In the case y_i is a vector that includes auxiliary variables with known totals, Hartley and Rao (1968) showed that the maximum likelihood estimator under simple random sampling is approximately equal to the traditional regression estimator of the total. This paper was the first to show how to incorporate known auxiliary population totals in a likelihood framework. For stratified random sampling, labels within strata are ignored but not strata labels because of known strata differences. The resulting maximum likelihood estimator is approximately equal to a pseudo-optimal linear regression estimator when auxiliary variables with known totals are available. The latter estimator has some good conditional design-based properties (see section 3.4). The focus of Hartley and Rao (1968) was on the estimation of a total, but the likelihood approach has much wider scope in sampling, including the estimation of distribution functions and quantiles and the construction of likelihood ratio based confidence intervals (see section 8.1). The Hartley-Rao non-parametric likelihood approach was discovered

independently twenty years later (Owen 1988) in the mainstream statistical inference under the name “empirical likelihood”. It has attracted a good deal of attention, including its application to various sampling problems. So in a sense the integration efforts with mainstream statistics were partially successful. Owen’s (2002) book presents a thorough account of empirical likelihood theory and its applications.

3.2 Model-Dependent Approach

The model-dependent approach to inference assumes that the population structure obeys a specified super-population model. The distribution induced by the assumed model provides inferences referring to the particular sample of units s that has been drawn. Such conditional inferences can be more relevant and appealing than repeated sampling inferences. But model-dependent strategies can perform poorly in large samples when the model is not correctly specified; even small deviations from the assumed model that are not easily detectable through model checking methods can cause serious problems. For example, consider the often-used ratio model when an auxiliary variable x with known total X is also measured in the sample:

$$y_i = \beta x_i + \varepsilon_i; i = 1, \dots, N \quad (2)$$

where the ε_i are independent random variables with zero mean and variance proportional to x_i . Assuming the model holds for the sample, that is, no sample selection bias, the best linear model-unbiased predictor of the total Y is given by the ratio estimator $(\bar{y}/\bar{x})X$ regardless of the sample design. This estimator is not design consistent unless the design is self-weighting, for example, stratified random sampling with proportional allocation. As a result, it can perform very poorly in large samples under non-self-weighting designs even if the deviations from the model are small. Hansen *et al.* (1983) demonstrated the poor performance under a repeated sampling set-up, using a stratified random sampling design with near optimal sample allocation (commonly used to handle highly skewed populations). Rao (1996) used the same design to demonstrate poor performance under a conditional framework relevant to the model-dependent approach (Royall and Cumberland 1981). Nevertheless, model-dependent approaches can play a vital role in small area estimation where the sample size in a small area (or domain) can be very small or even zero; see section 7.

Brewer (1963) proposed the model-dependent approach in the context of the ratio model (2). Royall (1970) and his collaborators made a systematic study of this approach. Valliant, Dorfman and Royall (2000) give a comprehensive account of the theory, including estimation of the (conditional) model variance of the estimator which varies with s .

For example, under the ratio model (2) the model variance depends on the sample mean \bar{x}_s . It is interesting to note that balanced sampling through purposive selection appears in the model-dependent approach in the context of protection against incorrect specification of the model (Royall and Herson 1973).

3.3 Model-Assisted Approach

The model-assisted approach attempts to combine the desirable features of design-based and model-dependent methods. It entertains only design-consistent estimators of the total Y that are also model unbiased under the assumed “working” model. For example, under the ratio model (2), a model-assisted estimator of Y for a specified probability sampling design is given by the ratio estimator $\hat{Y}_r = (\hat{Y}_{NHT} / \hat{X}_{NHT})X$ which is design consistent regardless of the assumed model. Hansen *et al.* (1983) used this estimator for their stratified design to demonstrate its superior performance over the model dependent estimator $(\bar{y}/\bar{x})X$. For variance estimation, the model-assisted approach uses estimators that are consistent for the design variance of the estimator and at the same time exactly or asymptotically model unbiased for the model variance. However, the inferences are design-based because the model is used only as a “working” model.

For the ratio estimator \hat{Y}_r , the variance estimator is given by

$$\text{var}(\hat{Y}_r) = (X / \hat{X}_{NHT})^2 v(e), \quad (3)$$

where in the operator notation $v(e)$ is obtained from $v(y)$ by changing y_i to the residuals $e_i = y_i - (\hat{Y}_{NHT} / \hat{X}_{NHT})x_i$. This variance estimator is asymptotically equivalent to a customary linearization variance estimator $v(e)$, but it reflects the fact that the information in the sample varies with \hat{X}_{NHT} : larger values lead to smaller variability and smaller values to larger variability. The resulting normal pivotal leads to valid model-dependent inferences under the assumed model (unlike the use of $v(e)$ in the pivotal) and at the same time protects against model deviations in the sense of providing asymptotically valid design-based inferences. Note that the pivotal is asymptotically equivalent to $\hat{Y}(\tilde{e}) / [\hat{Y}(\tilde{e})]^{1/2}$ with $\tilde{e}_i = y_i - (Y/X)x_i$. If the deviations from the model are not large, then the skewness in the residuals \tilde{e}_i will be small even if y_i and x_i are highly skewed, and normal confidence intervals will perform well. On the other hand, for highly skewed populations, the normal intervals based on \hat{Y}_{NHT} and its standard error may perform poorly under repeated sampling even for fairly large samples because the pivotal depends on the skewness of the y_i . Therefore, the population structure does matter in design-based inferences contrary to the claims of Neyman (1934), Hansen *et al.* (1983) and others. Rao, Jocelyn and Hidirolou (2003) considered the simple linear regression

estimator under two-phase simple random sampling with only x observed in the first phase. They demonstrated that the coverage performance of the associated normal intervals can be poor even for moderately large second phase samples if the true underlying model that generated the population deviated significantly from the linear regression model (for example, a quadratic regression of y on x) and the skewness of x is large. In this case, the first phase x -values are observed, and a proper model-assisted approach would use a multiple linear regression estimator with x and $z = x^2$ as the auxiliary variables. Note that for single phase sampling such a model-assisted estimator cannot be implemented if only the total X is known since the estimator depends on the population total of z .

Särndal, Swenson and Wretman (1992) provide a comprehensive account of the model-assisted approach to estimating the total Y of a variable y under the working linear regression model

$$y_i = x_i' \beta + \varepsilon_i; i = 1, \dots, N \quad (4)$$

with mean zero, uncorrelated errors ε_i and model variance $V_m(\varepsilon_i) = \sigma^2 q_i = \sigma_i^2$ where the q_i are known constants and the x -vectors have known totals X (the population values x_1, \dots, x_N may not be known). Under this set-up, the model-assisted approach leads to the generalized regression (GREG) estimator with a closed-form expression

$$\hat{Y}_{gr} = \hat{Y}_{NHT} + \hat{B}'(X - \hat{X}_{NHT}) =: \sum_{i \in S} w_i(s) y_i, \quad (5)$$

where

$$\hat{B} = \hat{T}^{-1} \left(\sum_s \pi_i^{-1} x_i y_i / q_i \right) \quad (6)$$

with $\hat{T} = \sum_s \pi_i^{-1} x_i x_i' / q_i$ is a weighted regression coefficient, and $w_i(s) = g_i(s) \pi_i^{-1}$ with $g_i(s) = 1 + (X - \hat{X}_{NHT})' \hat{T}^{-1} x_i / q_i$, known as “ g -weights”. Note that the GREG estimator (5) can also be written as $\sum_{i \in U} \hat{y}_i + \hat{E}_{NHT}$, where $\hat{y}_i = x_i' \hat{B}$ is the predictor of y_i under the working model and \hat{E}_{NHT} is the NHT estimator of the total prediction error $E = \sum_{i \in U} e_i$ with $e_i = y_i - \hat{y}_i$. This representation shows the role of the working model in the model-assisted approach. The GREG estimator (5) is design-consistent as well as model-unbiased under the working model (4). Moreover, it is nearly “optimal” in the sense of minimizing the asymptotic anticipated MSE (model expectation of the design MSE) under the working model, provided the inclusion probability, π_i , is proportional to the model standard deviation σ_i . However, in surveys with multiple variables of interest, the model variance may vary across variables. Because one must use a general-purpose design such as the design with inclusion probabilities proportional to sizes, the optimality result no longer holds, even if the same vector x_i is used for all the variables y_i in the working model.

The GREG estimator simplifies to the ‘projection’ estimator $X' \hat{B} = \sum_s w_i(s) y_i$ with $g_i(s) = X' \hat{T}^{-1} x_i / q_i$ if the model variance σ_i^2 is proportional to $\lambda' x_i$ for some λ . The ratio estimator is obtained as a special case of the projection estimator by letting $q_i = x_i$, leading to $g_i(s) = X / \hat{X}_{HT}$. Note that the GREG estimator (5) requires only the population totals X and not necessarily the individual population values x_i . This is very useful because the auxiliary population totals are often ascertained from external sources such as demographic projections of age and sex counts. Also, it ensures consistency with the known totals X in the sense of $\sum_s w_i(s) x_i = X$. Because of this property, GREG is also a calibration estimator.

Suppose there are p variables of interest, say $y^{(1)}, \dots, y^{(p)}$, and we want to use the model-assisted approach to estimate the corresponding population totals $Y^{(1)}, \dots, Y^{(p)}$. Also, suppose that the working model for $y^{(j)}$ is of the form (4) but requires possibly different x -vector $x^{(j)}$ with known total $X^{(j)}$ for each $j = 1, \dots, p$:

$$y_i^{(j)} = x_i^{(j)'} \beta^{(j)} + \varepsilon_i^{(j)}, i = 1, \dots, N. \quad (7)$$

In this case, the g -weights depend on j and in turn the final weights $w_i(s)$ also depend on j . In practice, it is often desirable to use a single set of final weights for all the p variables to ensure internal consistency of figures when aggregated over different variables. This property can be achieved only by enlarging the x -vector in the model (7) to accommodate all the variables $y^{(j)}$, say \tilde{x} with known total \tilde{X} and then using the working model

$$y_i^{(j)} = \tilde{x}_i' \beta^{(j)} + \varepsilon_i^{(j)}, i = 1, \dots, N. \quad (8)$$

However, the resulting weighted regression coefficients could become unstable due to possible multicollinearity in the enlarged set of auxiliary variables. As a result, the GREG estimator of $Y^{(j)}$ under model (8) is less efficient compared to the GREG estimator under model (7). Moreover, some of the resulting final weights, say $\tilde{w}_i(s)$, may not satisfy range restrictions by taking either values smaller than 1 (including negative values) or very large positive values. A possible solution to handle this problem is to use a generalized ridge regression estimator of $Y^{(j)}$ that is model-assisted under the enlarged model (Chambers 1996; Rao and Singh 1997).

For variance estimation, the model-assisted approach attempts to use design-consistent variance estimators that are also model-unbiased (at least for large samples) for the conditional model variance of the GREG estimator. Denoting the variance estimator of the NHT estimator of Y by $v(y)$ in an operator notation, a simple Taylor linearization variance estimator satisfying the above property is given by $v(ge)$, where $v(ge)$ is obtained by changing y_i to $g_i(s) e_i$ in the formula for $v(y)$; see Hidioglou, Fuller

and Hickman (1976) and Särndal, Swenson and Wretman (1989).

In the above discussion, we have assumed a working linear regression model for all the variables $y^{(j)}$. But in practice a linear regression model may not provide a good fit for some of the y -variables of interest, for example, a binary variable. In the latter case, logistic regression provides a suitable working model. A general working model that covers logistic regression is of the form $E_m(y_i) = h(x_i'\beta) = \mu_i$, where $h(\cdot)$ could be non-linear; model (5) is a special case with $h(a) = a$. A model-assisted estimator of the total under the general working model is the difference estimator $\hat{Y}_{\text{NHT}} + \sum_U \hat{\mu}_i - \sum_s \pi_i^{-1} \hat{\mu}_i$, where $\hat{\mu}_i = h(x_i'\hat{\beta})$ and $\hat{\beta}$ is an estimator of the model parameter β . It reduces to the GREG estimator (5) if $h(a) = a$. This difference estimator is nearly optimal if the inclusion probability π_i is proportional to σ_i , where σ_i^2 denotes the model variance, $V_m(y_i)$.

GREG estimators have become popular among users because many of the commonly used estimators may be obtained as special cases of (5) by suitable specifications of x_i and q_i . A Generalized Estimation System (GES) based on GREG has been developed at Statistics Canada.

Kott (2005) has proposed an alternative paradigm inference, called the randomization-assisted model-based approach, which attempts to focus on model-based inference assisted by randomization (or repeated sampling). The definition of anticipated variance is reversed to the randomization-expected model variance of an estimator, but it is identical to the customary anticipated variance when the working model holds for the sample, as assumed in the paper. As a result, the choices of estimator and variance estimator are often similar to those under the model-assisted approach. However, Kott argues that the motivation is clearer and “the approach proposed here for variance estimation leads to logically coherent treatment of finite population and small-sample adjustments when needed”.

3.4 Conditional Design-Based Approach

A conditional design-based approach has also been proposed. This approach attempts to combine the conditional features of the model-dependent approach with the model-free features of the design-based approach. It allows us to restrict the reference set of samples to a “relevant” subset of all possible samples specified by the design. Conditionally valid inferences are obtained in the sense that the conditional bias ratio (*i.e.*, the ratio of conditional bias to conditional standard error) goes to zero as the sample size increases. Approximately $100(1 - \alpha)\%$ of the realized confidence intervals in repeated sampling from the conditional set will contain the unknown total Y .

Holt and Smith (1979) provide compelling arguments in favour of conditional design based inference, even though the discussion was confined to one-way post-stratification of a simple random sample in which case it is natural to make inferences conditional on the realized strata sample sizes. Rao (1992, 1994) and Casady and Valliant (1993) studied conditional inference when only the auxiliary total X is known from external sources. In the latter case, conditioning on the NHT estimator \hat{X}_{NHT} may be reasonable because it is “approximately” an ancillary statistic when X is known and the difference $\hat{X}_{\text{NHT}} - X$ provides a measure of imbalance in the realized sample. Conditioning on \hat{X}_{NHT} leads to the “optimal” linear regression estimator which has the same form as the GREG estimator (5) with \hat{B} given by (6) replaced by the estimated optimal value \hat{B}_{opt} of the regression coefficient which involves the estimated covariance of \hat{Y}_{NHT} and \hat{X}_{NHT} and the estimated variance of \hat{X}_{NHT} . This optimal estimator leads to conditionally valid design-based inferences and model-unbiased under the working model (4). It is also a calibration estimator depending only on the total X and it can be expressed as $\sum_{i \in s} \tilde{w}_i(s) y_i$ with weights $\tilde{w}_i(s) = d_i \tilde{g}_i(s)$ and the calibration factor $\tilde{g}_i(s)$ depending only on the total X and the sample x -values. It works well for stratified random sampling (commonly used in establishment surveys). However, \hat{B}_{opt} can become unstable in the case of stratified multistage sampling unless the number of sample clusters minus the number of strata is fairly large. The GREG estimator does not require the latter condition but it can perform poorly in terms of conditional bias ratio and conditional coverage rates, as shown by Rao (1996). The unbiased NHT estimator can be very bad conditionally unless the design ensures that the measure of imbalance as defined above is small. For example, in the Hansen *et al.* (1983) design based on efficient x -stratification, the imbalance is small and the NHT estimator indeed performed well conditionally.

Tillé (1998) proposed an NHT estimator of the total Y based on approximate conditional inclusion probabilities given \hat{X}_{NHT} . His method also leads to conditionally valid inferences, but the estimator is not calibrated to X unlike the “optimal” linear regression estimator. Park and Fuller (2005) proposed a calibrated GREG version based on Tillé’s estimator which leads to non-negative weights more often than GREG.

I believe practitioners should pay more attention to conditional aspects of design-based inference and seriously consider the new methods that have been proposed.

Kalton (2002) has given compelling arguments for favoring design-based approaches (possibly model-assisted and/or conditional) for inference on finite population descriptive parameters. Smith (1994) named design-based inference as “procedural inference” and argued that

procedural inference is the correct approach for surveys in the public domain. We refer the reader to Smith (1976) and Rao and Bellhouse (1990) for reviews of inferential issues in sample survey theory.

4. Calibration Estimators

Calibration weights $w_i(s)$ that ensure consistency with user-specified auxiliary totals X are obtained by adjusting the design weights $d_i = \pi_i^{-1}$ to satisfy the benchmark constraints $\sum_{i \in s} w_i(s) x_i = X$. Estimators that use calibration weights are called calibration estimators and they use a single set of weights $\{w_i(s)\}$ for all the variables of interest. We have noted in section 3.4 that the model-assisted GREG estimator is a calibration estimator, but a calibration estimator may not be model-assisted in the sense that it could be model-biased under a working model (4) unless the x -variables in the model exactly match the variables corresponding to the user-specified totals. For example, suppose the working model suggested by the data is a quadratic in a scalar variable x while the user-specified total is only its total X . The resulting calibration estimator can perform poorly even in fairly large samples, as noted in section 3.3, unlike the model-assisted GREG estimator based on the working quadratic model that requires the population total of the quadratic variables x_i^2 in addition to X .

Post-stratification has been extensively used in practice to ensure consistency with known cell counts corresponding to a post-stratification variable, for example counts in different age groups ascertained from external sources such as demographic projections. The resulting post-stratified estimator is a calibration estimator. Calibration estimators that ensure consistency with known marginal counts of two or more post-stratification variables have also been employed in practice; in particular raking ratio estimators that are obtained by benchmarking to the marginal counts in turn until convergence is approximately achieved, typically in four or less iterations. Raking ratio weights $w_i(s)$ are always positive. In the past, Statistics Canada used raking ratio estimators in the Canadian Census to ensure consistency of 2B-item estimators with known 2A-item counts. In the context of the Canadian Census, Brackstone and Rao (1979) studied the efficiency of raking ratio estimators and also derived Taylor linearization variance estimators when the number of iterations is four or less. Raking ratio estimators have also been employed in the U.S. Current Population Survey (CPS). It may be noted that the method of adjusting cell counts to given marginal counts in a two-way table was originally proposed in the landmark paper by Deming and Stephan (1940).

Unified approaches to calibration, based on minimizing a suitable distance measure between calibration weights and design weights subject to benchmark constraints, have attracted the attention of users due to their ability to accommodate arbitrary number of user-specified benchmark constraints, for example, calibration to the marginal counts of several post-stratification variables. Calibration software is also readily available, including GES (Statistics Canada), LIN WEIGHT (Statistics Netherlands), CALMAR (INSEE, France) and CLAN97 (Statistics Sweden).

A chi-squared distance, $\sum_{i \in s} q_i (d_i - w_i)^2 / d_i$, leads to the GREG estimator (5), where the x -vector corresponds to the user-specified benchmark constraints (BC) and $w_i(s)$ is denoted as w_i for simplicity (Huang and Fuller 1978; Deville and Särndal 1992). However, the resulting calibration weights may not satisfy desirable range restrictions (RR), for example some weights may be negative or too large especially when the number of constraints is large and the variability of the design weights is large. Huang and Fuller (1978) proposed a scaled modified chi-squared distance measure and obtained the calibration weights through an iterative solution that satisfies BC at each iteration. However, a solution that satisfies BC and RR may not exist. Another method, called shrinkage minimization (Singh and Mohl 1996) has the same difficulty. Quadratic programming methods that minimize the chi-squared distance subject to both BC and RR have also been proposed (Hussain 1969) but the feasible set of solutions satisfying both BC and RR can be empty. Alternative methods propose to change the distance function (Deville and Särndal 1992) or drop some of the BC (Bankier, Rathwell and Majkowski 1992). For example, an information distance of the form $\sum_{i \in s} q_i \{w_i \log(w_i / d_i) - w_i + d_i\}$ gives raking ratio estimators with non-negative weights w_i , but some of the weights can be excessively large. "Ridge" weights obtained by minimizing a penalized chi-squared distance have also been proposed (Chambers 1996), but no guarantee that either BC or RR are satisfied, although the weights are more stable than the GREG weights. Rao and Singh (1997) proposed a "ridge shrinkage" iterative method that ensures convergence for a specified number of iterations by using a built-in tolerance specification to relax some BC while satisfying RR. Chen, Sitter and Wu (2002) proposed a similar method.

GREG calibration weights have been used in the Canadian Labour Force Survey and more recently it has been extended to accommodate composite estimators that make use of sample information in previous months, as noted in section 2 (Fuller and Rao 2001; Gambino, Kennedy and Singh 2001; Singh, Kennedy and Wu 2001). GREG-type calibration estimators have also been used for the integration of two or more independent surveys from the

same population. Such estimators ensure consistency between the surveys, in the sense that the estimates from the two surveys for common variables are identical, as well as benchmarking to known population totals (Renssen and Nieuwenbroek 1997; Singh and Wu 1996; Merkouris 2004). For the 2001 Canadian Census, Bankier (2003) studied calibration weights corresponding to the “optimal” linear regression estimator (section 3.3) under stratified random sampling. He showed that the “optimal” calibration method performed better than the GREG calibration used in the previous census, in the sense of allowing more BC to be retained while at the same time allowing the calibration weights to be at least one. The “optimal” calibration weights can be obtained from GES software by including the known strata sizes in the BC and defining the tuning constant q_i suitably. Note that the “optimal” calibration estimator also has desirable conditional design properties (section 3.4). Weighting for the 2001 Canadian census switched from projection GREG (used in the 1996 census) to “optimal” linear regression.

Demnati and Rao (2004) derived Taylor linearization variance estimators for a general class of calibration estimators with weights $w_i = d_i F(x_i' \hat{\lambda})$, where the LaGrange multiplier $\hat{\lambda}$ is determined by solving the calibration constraints. The choice $F(a) = 1 + a$ gives GREG weights and $F(a) = e^a$ leads to raking ratio weights. In the special case of GREG weights, the variance estimator reduces to $v(\text{ge})$ given in section 3.3.

We refer the reader to the Waksberg award paper of Fuller (Fuller 2002) for an excellent overview and appraisal of regression estimation in survey sampling, including calibration estimation.

5. Unequal Probability Sampling Without Replacement

We have noted in section 2 that PPS sampling of PSUs within strata in large-scale surveys was practically motivated by the desire to achieve approximately equal workloads. PPS sampling also achieves significant variance reduction by controlling on the variability arising from unequal PSU sizes without actually stratifying by size. PSUs are typically sampled without replacement such that the PSU inclusion probability, π_i , is proportional to PSU size measure x_i . For example, systematic PPS sampling, with or without initial randomization of the PSU labels, is an inclusion probability proportional to size (IPPS) design (also called π PS design) that has been used in many complex surveys, including the Canadian LFS. The estimator of a total associated with an IPPS design is the NHT estimator.

Development of suitable (IPPS, NHT) strategies raises theoretically challenging problems, including the evaluation

of exact joint inclusion probabilities, π_{ij} , or accurate approximations to π_{ij} requiring only the individual π_i s, that are needed in getting unbiased or nearly unbiased variance estimator. My own 1961 Ph.D. thesis at Iowa State University addressed the latter problem. Several solutions, requiring sophisticated theoretical tools, have been published since then by talented mathematical statisticians. However, this theoretical work is often classified as “theory without application” because it is customary practice to treat the PSUs as if sampled with replacement since that leads to great simplification. The variance estimator is simply obtained from the estimated PSU totals and, in fact, this assumption is the basis for re-sampling methods (section 6). This variance estimator can lead to substantial over-estimation unless the overall PSU sampling fraction is small. The latter may be true in many large-scale surveys. In the following paragraphs, I will try to demonstrate that the theoretical work on (IPPS, NHT) strategies as well as some non-IPPS designs have wide practical applicability.

First, I will focus on (IPPS, NHT) strategies. In Sweden and some other countries in Europe, stratified single-stage sampling is often used because of the availability of list frames and IPPS designs are attractive options, but sampling fractions are often large. For example, Rosén (1991) notes that Statistics Sweden’s Labour Force Barometer samples some 100 different populations using systematic PPS sampling and that the sampling rates can exceed 50%. Aires and Rosén (2005) studied Pareto π PS sampling for Swedish surveys. This method has attractive properties, including fixed sample size, simple sample selection, good estimation precision and consistent variance estimation regardless of sampling rates. It also allows sample coordination through permanent random numbers (PRN) as in Poisson sampling, but the latter method leads to variable sample size. Because of these merits, Pareto π PS has been implemented in a number of Statistics Sweden surveys, notably in price index surveys. Ohlsson (1995) described PRN techniques that are commonly used in practice.

The method of Rao-Sampford (see Brewer and Hanif 1983, page 28) leads to exact IPPS designs and non-negative unbiased variance estimators for arbitrary fixed sample sizes. It has been implemented in the new version of SAS. Stehman and Overton (1994) note that variable probability structure arises naturally in environmental surveys rather than being selected just for enhanced efficiency, and that the π_i s are only known for the units i in the sample s . By treating the sample design as randomized systematic PPS, Stehman and Overton obtained approximations to the π_{ij} s that depend only $\pi_i, i \in s$, unlike the original approximations of Hartley and Rao (1962) that require the sum of squares of all the π_i s in the population. In the Stehman and Overton applications, the sampling rates are

substantial enough to warrant the evaluation of the joint inclusion probabilities.

I will now turn to non-IPPS designs using estimators different from the NHT estimator that ensure zero variance when y is exactly proportional to x . The random group method of Rao, Hartley and Cochran (1962) permits a simple non-negative variance estimator for any fixed sample size and yet compares favorably to (IPPS, NHT) strategies in terms of efficiency and is always more efficient than the PPS with replacement strategy. Schabenberger and Gregoire (1994) noted that (IPPS, NHT) strategies have not enjoyed much application in forestry because of difficulty in implementation and recommended the Rao-Hartley-Cochran strategy in view of its remarkable simplicity and good efficiency properties. It is interesting to note that this strategy has been used in the Canadian LFS on the basis of its suitability for switching to new size measures, using the Keyfitz method within each random group. On the other hand, (IPPS, NHT) strategies are not readily suitable for this purpose (Fellegi 1966). I understand that the Rao-Hartley-Cochran strategy is often used in audit sampling and other accounting applications.

Murthy (1957) used a non-IPPS design based on drawing successive units with probabilities $p_i, p_j/(1-p_i), p_k/(1-p_i-p_j)$ and so on, and the following estimator:

$$\hat{Y}_M = \sum_{i \in s} y_i \frac{p(s|i)}{p(s)}, \quad (9)$$

where $p(s|i)$ is the conditional probability of obtaining the sample s given that unit i was selected first. He also provided a non-negative variance estimator requiring the conditional probabilities, $p(s|i, j)$, of obtaining s given i and j are selected in the first two draws. This method did not receive practical attention for several years due to computational complexity, but more recently it has been applied in unexpected areas, including oil discovery (Andreatta and Kaufmann 1986) and sequential sampling including inverse sampling and some adaptive sampling schemes (Salehi and Seber 1997). It may be noted that adaptive sampling has received a lot of attention in recent years because of its potential as an efficient sampling method for estimating totals or means of rare populations (Thompson and Seber 1996). In the oil discovery application, the successive sampling scheme is a characterization of discovery and the order in which fields are discovered is governed by sampling proportional to field size and without replacement, following the industry folklore "on the average, the big fields are found first". Here $p_i = y_i/Y$ and the total oil reserve Y is assumed to be known from geological considerations. In this application, geologists are interested in the size distribution of all fields in the basin and when a basin is partially explored the sample is composed

of magnitudes y_i of discovered deposits. The size distribution function $F(a)$ can be estimated by using Murthy's estimator (9) with y_i replaced by the indicator variable $I(y_i \leq a)$. The computation of $p(s|i)$ and $p(s)$, however, is formidable even for moderate sample sizes. To overcome this computational difficulty, Andreatta and Kaufman (1986) used integral representations of these quantities to develop asymptotic expansions of Murthy's estimator, the first few terms of which are easily computable. Similarly, they obtain computable approximations to Murthy's variance estimator. Note that the NHT estimator of $F(a)$ is not feasible here because the inclusion probabilities are functions of all the y -values in the population.

The above discussion is intended to demonstrate that a particular theory can have applications in diverse practical areas even if it is not needed in a particular situation, such as large-scale surveys with negligible first stage sampling fractions. Also it shows that unequal probability sampling designs play a vital role in survey sampling, despite Särndal's (1996) contention that simpler designs, such as stratified SRS and stratified Bernoulli sampling, together with GREG estimators should replace strategies based on unequal probability sampling without replacement.

6. Analysis of Survey Data and Resampling Methods

Standard methods of data analysis are generally based on the assumption of simple random sampling, although some software packages do take account of survey weights and provide correct point estimates. However, application of standard methods to survey data, ignoring the design effect due to clustering and unequal probabilities of selection, can lead to erroneous inferences even for large samples. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated, type I error rates of tests of hypotheses can be much bigger than the nominal levels, and standard model diagnostics, such as residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of those problems and emphasized the need for new methods that take proper account of the complexity of data derived from large-scale surveys. Fuller (1975) developed asymptotically valid methods for linear regression analysis, based on Taylor linearization variance estimators. Rapid progress has been made over the past 20 years or so in developing suitable methods. Resampling methods play a vital role in developing methods that take account of survey design in the analysis of data. All one needs is a data file containing the observed data, the final survey weights and the corresponding final weights for each pseudo-replicate generated by the re-sampling method. Software packages that take account of survey weights in

the point estimation of parameters of interest can then be used to calculate the correct estimators and standard errors, as demonstrated below. As a result, re-sampling methods of inference have attracted the attention of users as they can perform the analyses themselves very easily using standard software packages. However, releasing public-use data files with replicate weights can lead to confidentiality issues, such as the identification of clusters from replicate weights. In fact, at present a challenge to theory is to develop suitable methods that can preserve confidentiality of the data. Lu, Brick and Sitter (2004) proposed grouping strata and then forming pseudo-replicates using the combined strata for variance estimation, thus limiting the risk of cluster identification from the resulting public-use data file. Grouping strata and/or PSUs within strata simplifies variance estimation by reducing the number of pseudo-replicates used in variance estimation compared to the commonly used delete-cluster jackknife discussed below. A method of inverse sampling to undo the complex survey data structure and yet provide protection against revealing cluster labels (Hinkins, Oh and Scheuren 1997; Rao, Scott and Benhin 2003) appears promising, but much work on inverse sampling methods remains to be done before it becomes attractive to the user.

Rao and Scott (1981, 1984) made a systematic study of the impact of survey design effect on standard chi-squared and likelihood ratio tests associated with a multi-way table of estimated counts or proportions. They showed that the test statistic is asymptotically distributed as a weighted sum of independent χ^2_1 variables, where the weights are the eigenvalues of a "generalized design effects" matrix. This general result shows that the survey design can have a substantial impact on the type I error rate. Rao and Scott proposed simple first-order corrections to the standard chi-squared statistics that can be computed from published tables that include estimates of design effects for cell estimates and their marginal totals, thus facilitating secondary analyses from published tables. They also derived second order corrections that are more accurate, but require the knowledge of a full estimated covariance matrix of the cell estimates, as in the case of familiar Wald tests. However, Wald tests can become highly unstable as the number of cells in a multi-way table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal levels, unlike the Rao-Scott second order corrections (Thomas and Rao 1987). The first and second order corrections are now known as Rao-Scott corrections and are given as default options in the new version of SAS. Roberts, Rao and Kumar (1987) developed Rao-Scott type corrections to tests for logistic regression analysis of estimated cell proportions associated with a binary response variable. They applied the methods

to a two-way table of employment rates from the Canadian LFS 1977 obtained by cross-classifying age and education groups. Bellhouse and Rao (2002) extended the work of Roberts *et al.* to the analysis of domain means using generalized linear models. They applied the methods to domain means from a Fiji Fertility Survey cross-classified by education and years since the woman's first marriage, where a domain mean is the mean number of children ever born for women of Indian race belonging to the domain.

Re-sampling methods in the context of large-scale surveys using stratified multi-stage designs have been studied extensively. For inference purposes, the sample PSUs are treated as if drawn with replacement within strata. This leads to over-estimation of variances but it is small if the overall PSU sampling fraction is negligible. Let $\hat{\theta}$ be the survey-weighted estimator of a "census" parameter of interest computed from the final weights w_i , and let the corresponding weights for each pseudo-replicate r generated by the re-sampling method be denoted by $w_i^{(r)}$. The estimator based on the pseudo-replicate weights $w_i^{(r)}$ is denoted as $\hat{\theta}^{(r)}$ for each $r = 1, \dots, R$. Then a re-sampling variance estimator of $\hat{\theta}$ is of the form

$$v(\hat{\theta}) = \sum_{r=1}^R c_r (\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})' \quad (10)$$

for specified coefficients c_r in (10) determined by the re-sampling method.

Commonly used re-sampling methods include (a) delete-cluster (delete-PSU) jackknife, (b) balanced repeated replication (BRR) particularly for $n_h = 2$ PSUs in each stratum h and (c) the Rao and Wu (1988) bootstrap. Jackknife pseudo-replicates are obtained by deleting each sample cluster $r = (hj)$ in turn, leading to jackknife design weights $d_i^{(r)}$ taking the value 0 if the sample unit i is in the deleted cluster, $n_h d_i / (n_h - 1)$ if i is not in the deleted cluster but in the same stratum, and unchanged if i is in a different stratum. The jackknife design weights are then adjusted for unit non-response and post-stratification, leading to the final jackknife weights $w_i^{(r)}$. The jackknife variance estimator is given by (10) with $c_r = (n_h - 1) / n_h$ for $r = (hj)$. The delete-cluster jackknife method has two possible disadvantages: (1) When the total number of sampled PSUs, $n = \sum n_h$, is very large, R is also very large because $R = n$. (2) It is not known if the delete-jackknife variance estimator is design-consistent in the case of non-smooth estimators $\hat{\theta}$, for example the survey-weighted estimator of the median. For simple random sampling, the jackknife is known to be inconsistent for the median or other quantiles. It would be theoretically challenging and practically relevant to find conditions for the consistency of the delete-cluster jackknife variance estimator of a non-smooth estimator $\hat{\theta}$.

BRR can handle non-smooth $\hat{\theta}$, but it is readily applicable only for the important special case of $n_h = 2$ PSUs per stratum. A minimal set of balanced half-samples can be constructed from an $R \times R$ Hadamard matrix by selecting H columns, excluding the column of +1's, where $H + 1 \leq R \leq H + 4$ (McCarthy 1969). The BRR design weights $d_i^{(r)}$ equal $2d_i$ or 0 according as whether or not i is in the half-sample. A modified BRR, due to Bob Fay, uses all the sampled units in each replicate unlike the BRR by defining the replicate design weights as $d_i^{(r)}(\epsilon) = (1 + \epsilon)d_i$ or $(1 - \epsilon)d_i$ according as whether or not i is in the half-sample, where $0 < \epsilon < 1$; a good choice of ϵ is $1/2$. The modified BRR weights are then adjusted for non-response and post-stratification to get the final weights $w_i^{(r)}(\epsilon)$ and the estimator $\hat{\theta}^{(r)}(\epsilon)$. The modified BRR variance estimator is given by (10) divided by ϵ^2 and $\hat{\theta}^{(r)}$ replaced by $\hat{\theta}^{(r)}(\epsilon)$; see Rao and Shao (1999). The modified BRR is particularly useful under independent re-imputation for missing item responses in each replicate because it can use the donors in the full sample to impute unlike the BRR that uses the donors only in the half-sample.

The Rao-Wu bootstrap is valid for arbitrary $n_h (\geq 2)$ unlike the BRR, and it can also handle non-smooth $\hat{\theta}$. Each bootstrap replicate is constructed by drawing a simple random sample of PSUs of size $n_h - 1$ from the n_h sample clusters, independently across the strata. The bootstrap design weights $d_i^{(r)}$ are given by $[n_h / (n_h - 1)]m_{hi}^{(r)}d_i$ if i is in stratum h and replicate r , where $m_{hi}^{(r)}$ is the number of times sampled PSU (hi) is selected, $\sum_i m_{hi}^{(r)} = n_h - 1$. The weights $d_i^{(r)}$ are then adjusted for unit non-response and post-stratification to get the final bootstrap weights and the estimator $\hat{\theta}^{(r)}$. Typically, $R = 500$ bootstrap replicates are used in the bootstrap variance estimator (10). Several recent surveys at Statistics Canada have adopted the bootstrap method for variance estimation because of the flexibility in the choice of R and wider applicability. Users of Statistics Canada survey micro data files seem to be very happy with the bootstrap method for analysis of data.

Early work on the jackknife and the BRR was largely empirical (e.g., Kish and Frankel 1974). Krewski and Rao (1981) formulated a formal asymptotic framework appropriate for stratified multi-stage sampling and established design consistency of the jackknife and BRR variance estimators when $\hat{\theta}$ can be expressed as a smooth function of estimated means. Several extensions of this basic work have been reported in the recent literature; see the book by Shao and Tu (1995, Chapter 6). Theoretical support for re-sampling methods is essential for their use in practice.

In the above discussion, I let $\hat{\theta}$ denote the estimator of a "census" parameter. The census parameter θ_C is often motivated by an underlying super-population model and the census is regarded as a sample generated by the model, leading to census estimating equations whose solution is

θ_C . The census estimating functions $U_C(\theta)$ are simply population totals of functions $u_i(\theta)$ with zero expectation under the assumed model, and the census estimating equations are given by $U_C(\theta) = 0$ (Godambe and Thompson 1986). Kish and Frankel (1974) argued that the census parameter makes sense even if the model is not correctly specified. For example, in the case of linear regression, the census regression coefficient could explain how much of the relationship between the response variable and the independent variables is accounted by a linear regression model. Noting that the census estimating functions are simply population totals, survey weighted estimators $\hat{U}(\theta)$ from the full sample and $\hat{U}^{(r)}(\theta)$ from each pseudo-replicate are obtained. The solutions of corresponding estimating equations $\hat{U}(\theta) = 0$ and $\hat{U}^{(r)}(\theta) = 0$ give $\hat{\theta}$ and $\hat{\theta}^{(r)}$ respectively. Note that the re-sampling variance estimators are designed to estimate the variance of $\hat{\theta}$ as an estimator of the census parameters but not the model parameters. Under certain conditions, the difference can be ignored but in general we have a two-phase sampling situation, where the census is the first phase sample from the super-population and the sample is a probability sample from the census population. Recently, some useful work has been done on two-phase variance estimation when the model parameters are the target parameters (Graubard and Korn 2002; Rubin-Bleuer and Schiopu-Kratina 2005), but more work is needed to address the difficulty in specifying the covariance structure of the model errors.

A difficulty with the bootstrap is that the solution $\hat{\theta}^{(r)}$ may not exist for some bootstrap replicates r (Binder, Kovacevic and Roberts 2004). Rao and Tausi (2004) used an estimating function (EF) bootstrap method that avoids the difficulty. In this method, we solve $\hat{U}(\theta) = \hat{U}^{(r)}(\hat{\theta})$ for θ using only one step of the Newton-Raphson iteration with $\hat{\theta}$ as the starting value. The resulting estimator $\tilde{\theta}^{(r)}$ is then used in (10) to get the EF bootstrap variance estimator of $\hat{\theta}$ which can be readily implemented from the data file providing replicate weights, using slight modifications of any software package that accounts for survey weights. It is interesting to note that the EF bootstrap variance estimator is equivalent to a Taylor linearization sandwich variance estimator that uses the bootstrap variance estimator of $\hat{U}(\theta)$ and the inverse of the observed information matrix (derivative of $-\hat{U}(\theta)$), both evaluated at $\theta = \hat{\theta}$ (Binder *et al.* 2004).

Taylor linearization methods provide asymptotically valid variance estimators for general sampling designs, unlike re-sampling methods, but they require a separate formula for each estimator $\hat{\theta}$. Binder (1983), Rao, Yung and Hidioglou (2002) and Demnati and Rao (2004) have provided unified linearization variance formulae for estimators defined as solutions to estimating equations.

Pfeffermann (1993) discussed the role of design weights in the analysis of survey data. If the population model holds for the sample (*i.e.*, if there is no sample selection bias), then model-based unweighted estimators will be more efficient than the weighted estimators and lead to valid inferences, especially for data with smaller sample sizes and larger variation in the weights. However, for typical data from large-scale surveys, the survey design is informative and the population model may not hold for the sample. As a result, the model-based estimators can be seriously biased and inferences can be erroneous. Pfeffermann and his colleagues initiated a new approach to inference under informative sampling; see Pfeffermann and Sverchokov (2003) for recent developments. This approach seems to provide more efficient inferences compared to the survey weighted approach, and it certainly deserves the attention of users of survey data. However, much work remains to be done, especially in handling data based on multi-stage sampling.

Excellent accounts of methods for analysis of complex survey data are given in Skinner, Holt and Smith (1989), Chambers and Skinner (2003) and Lehtonen and Pahkinen (2004).

7. Small Area Estimation

Previous sections of this paper have focussed on traditional methods that use direct domain estimators based on domain-specific sample observations along with auxiliary population information. Such methods, however, may not provide reliable inferences when the domain sample sizes are very small or even zero for some domains. Domains or sub-populations with small or zero sample sizes are called small areas in the literature. Demand for reliable small area statistics has greatly increased in recent years because of the growing use of small area statistics in formulating policies and programs, allocation of funds and regional planning. Clearly, it is seldom possible to have a large enough overall sample size to support reliable direct estimates for all domains of interest. Also, in practice, it is not possible to anticipate all uses of survey data and “the client will always require more than is specified at the design stage” (Fuller 1999, page 344). In making estimates for small areas with adequate level of precision, it is often necessary to use “indirect” estimators that borrow information from related domains through auxiliary information, such as census and current administrative data, to increase the “effective” sample size within the small areas.

It is now generally recognized that explicit models linking the small areas through auxiliary information and accounting for residual between – area variation through random small area effects are needed in developing indirect estimators. Success of such model-based methods heavily

depends on the availability of good auxiliary information and thorough validation of models through internal and external evaluations. Many of the random effects methods used in mainstream statistical theory are relevant to small area estimation, including empirical best (or Bayes), empirical best linear unbiased prediction and hierarchical Bayes based on prior distributions on the model parameters. A comprehensive account of such methods is given in Rao (2003). Practical relevance and theoretical interest of small area estimation have attracted the attention of many researchers, leading to important advances in point and mean squared error estimation. The “new” methods have been applied successfully worldwide to a variety of small area problems. Model-based methods have been recently used to produce county and school district estimates of poor school-age children in the U.S.A. The U.S. Department of Education allocates annually over \$7 billion of funds to counties on the basis of model-based county estimates. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children. We refer to Rao (2003, example 7.1.2) for details of this application. In the United Kingdom, the Office for National Statistics established a Small Area Estimation Project to develop model-based estimates at the level of political wards (roughly 2,000 households). The practice and estimation methods of U.S. federal statistical programs that use indirect estimators to produce published estimates are documented in Schaible (1996). Singh, Gambino and Mantel (1994) and Brackstone (2002) discuss some practical issues and strategies for small area statistics.

Small area estimation is a striking example of the interplay between theory and practice. The theoretical advances are impressive, but many practical issues need further attention of theory. Such issues include: (a) Benchmarking model-based estimators to agree with reliable direct estimators at large area levels. (b) Developing and validating suitable linking models and addressing issues such as errors in variables, incorrect specification of the linking model and omitted variables. (c) Development of methods that satisfy multiple goals: good area-specific estimates, good rank properties and good histogram for small areas.

8. Some Theory Deserving Attention of Practice and Vice Versa

In this section, I will briefly mention some examples of important theory that exists but not widely used in practice.

8.1 Empirical Likelihood Inference

Traditional sampling theory largely focused on point estimation and associated standard errors, appealing to normal approximations for confidence intervals on parameters

of interest. In mainstream statistics, the empirical likelihood (EL) approach (Owen 1988) has attracted a lot of attention due to several desirable properties. It provides a non-parametric likelihood, leading to EL ratio confidence intervals similar to the parametric likelihood ratio intervals. The shape and orientation of EL intervals are determined entirely by the data, and the intervals are range preserving and transformation respecting, and are particularly useful in providing balanced tail error rates, unlike the symmetric normal theory intervals. As noted in section 3.1, the EL approach was in fact first introduced in the sample survey context by Hartley and Rao (1968), but their focus was on inferential issues related to point estimation. Chen, Chen and Rao (2003) obtained EL intervals on the population mean under simple random and stratified random sampling for populations containing many zeros. Such populations are encountered in audit sampling, where y denotes the amount of money owed to the government and the mean \bar{Y} is the average amount of excessive claims. Previous work on audit sampling used parametric likelihood ratio intervals based on parametric mixture distributions for the variable y . Such intervals perform better than the standard normal theory intervals, but EL intervals perform better under deviations from the assumed mixture model, by providing non-coverage rate below the lower bound closer to the nominal error rate and also larger lower bound. For general designs, Wu and Rao (2004) used a pseudo-empirical likelihood (Chen and Sitter 1999) to obtain adjusted pseudo-EL intervals on the mean and the distribution function that account for the design features, and showed that the intervals provide more balanced tail error rates than the normal theory intervals. The EL method also provides a systematic approach to calibration estimation and integration of surveys. We refer the reader to the review papers by Rao (2004) and Wu and Rao (2005).

Further refinements and extensions remain to be done, particularly on the pseudo-empirical likelihood, but the EL theory in the survey context deserves the attention of practice.

8.2 Exploratory Analyses of Survey Data

In section 6 we discussed methods for confirmatory analysis of survey data taking the design into account, such as point estimation of model (or census) parameters and associated standard errors and formal tests of hypotheses. Graphical displays and exploratory data analyses of survey data are also very useful. Such methods have been extensively developed in the mainstream literature. Only recently, some extensions of these modern methods are reported in the survey literature and deserve the attention of practice. I will briefly mention some of those developments. First, non-parametric kernel density estimates are commonly used to

display the shape of a data set without relying on parametric models. They can also be used to compare different sub-populations.

Bellhouse and Stafford (1999) provided kernel density estimators that take account of the survey design and studied their properties and applied the methods to data from the Ontario Health Survey. Buskirk and Lohr (2005) studied asymptotic and finite sample properties of kernel density estimators and obtained confidence bands. They applied the methods to data from the US National Crime Victimization Survey and the US National Health and Nutrition Examination Survey.

Secondly, Bellhouse and Stafford (2001) developed local polynomial regression methods, taking design into account, that can be used to study the relationship between a response variable and predictor variables, without making strong parametric model assumptions. The resulting graphical displays are useful in understanding the relationships and also for comparing different sub-populations. Bellhouse and Stafford (2001) illustrated local polynomial regression on the Ontario Health Survey data; for example, the relationship between body mass index of females and age. Bellhouse, Chipman and Stafford (2004) studied additive models for survey data via penalized least squares method to handle more than one predictor variable, and illustrated the methods on the Ontario Health Survey data. This approach has many advantages in terms of graphical display, estimation, testing and selection of “smoothing” parameters for fitting the models.

8.3 Measurement Errors

Typically, measurement errors are assumed to be additive with zero means. As a result, usual estimators of totals and means remain unbiased or consistent. However, this nice feature may not hold for more complex parameters such as distribution functions, quantiles and regression coefficients. In the latter case, the usual estimators will be biased, even for large samples, and hence can lead to erroneous inferences (Fuller 1995). It is possible to obtain bias-adjusted estimators if estimates of measurement error variances are available. The latter may be obtained by allocating resources at the design stage to make repeated observations on a sub-sample. Fuller (1975, 1995) has been a champion of proper methods in the presence of measurement errors and the bias-adjusted methods deserve the attention of practice.

Hartley and Rao (1978) and Hartley and Biemer (1978) provided interviewer and coder assignment conditions that permit the estimation of sampling and response variances for the mean or total from current surveys. Unfortunately, current surveys are often not designed to satisfy those conditions and even if they do the required information on

interviewer and coder assignments is seldom available at the estimation stage.

Linear components of variance models are often used to estimate interviewer variability. Such models are appropriate for continuous responses, but not for binary responses. The linear model approach for binary responses can result in underestimating the intra-interviewer correlations. Scott and Davis (2001) proposed multi-level models for binary responses to estimate interviewer variability. Given that responses are often binary in many surveys, practice should pay attention to such models for proper analyses of survey data with binary responses.

8.4 Imputation for Missing Survey Data

Imputation is commonly used in practice to fill in missing item values. It ensures that the results obtained from different analyses of the completed data set are consistent with one another by using the same survey weight for all items. Marginal imputation methods, such as ratio, nearest neighbor and random donor within imputation classes are used by many statistical agencies. Unfortunately, the imputed values are often treated as if they were true values and then used to compute estimates and variance estimates. The imputed point estimates of marginal parameters are generally valid under an assumed response mechanism or imputation model. But the “naïve” variance estimators can lead to erroneous inferences even for large samples; in particular, serious underestimation of the variance of the imputed estimator because the additional variability due to estimating the missing values is not taken into account. Advocates of Rubin’s (1987) multiple imputation claim that the multiple imputation variance estimator can fix this problem because a between imputed estimators sum of squares is added to the average of naïve variance estimators resulting from the multiple imputations. Unfortunately, there are some difficulties associated with multiple imputation variance estimators, as discussed by Kott (1995), Fay (1996), Binder and Sun (1996), Wang and Robins (1998), Kim, Brick, Fuller and Kalton (2004) and others. Moreover, single imputation is often preferred due to operational and cost considerations. Some impressive advances have been made in recent years on making efficient and asymptotically valid inferences from singly imputed data sets. We refer the reader to review papers by Shao (2002) and Rao (2000, 2005) for methods of variance estimation under single imputation. Kim and Fuller (2004) studied fractional imputation using more than one randomly imputed value and showed that it also leads to asymptotically valid inferences; see also Kalton and Kish (1984) and Fay (1996). An advantage of fractional imputation is that it reduces the imputation variance relative to single imputation using one randomly imputed value. The above methods of variance estimation deserve the attention of practice.

8.5 Multiple Frame Surveys

Multiple frame surveys employ two or more overlapping frames that can cover the target population. Hartley (1962) studied the special case of a complete frame B and an incomplete frame A and simple random sampling independently from both frames. He showed that an “optimal” dual frame estimator can lead to large gains in efficiency for the same cost over the single complete frame estimator, provided the cost per unit for frame A is significantly smaller than the cost per unit for frame B . Multiple frame surveys are particularly suited for sampling rare or hard-to-reach populations, such as homeless populations and persons with AIDS, when incomplete list frames contain high proportions of individuals from the target population. Hartley’s (1974) landmark paper derived “optimal” dual frame estimators for general sampling designs and possibly different observational units in the two frames. Fuller and Burneister (1972) proposed improved “optimal” estimators. However, the optimal estimators use different sets of weights for each item y , which is not desirable in practice. Skinner and Rao (1996) derived pseudo-ML (PML) estimators for dual frame surveys that use the same set of weights for all items y , similar to “single frame” estimators (Kalton and Anderson 1986), and maintain efficiency. Lohr and Rao (2005) developed a unified theory for the multiple frames setting with two or more frames, by extending the optimal, pseudo-ML and single frame estimators. Lohr and Rao (2000, 2005) obtained asymptotically valid jackknife variance estimators. Those general results deserve the attention of practice when dealing with two or more frames. Dual frame telephone surveys based on cell phone and landline phone frames need the attention of theory because it is unclear how to weight in the cell phone survey: some families share a cell phone and others have a cell phone for each person.

8.6 Indirect Sampling

The method of indirect sampling can be used when the frame for a target population U^B is not available but the frame for another population U^A , linked to U^B , is employed to draw a probability sample. The links between the two populations are used to develop suitable weights that can provide unbiased estimators and variance estimators. Lavallée (2002) developed a unified method, called Generalized Weight Sharing, (GWS), that covers several known methods: the weight sharing method of Ernst (1989) for cross sectional estimation from longitudinal household surveys, network sampling and multiplicity estimation (Sirken 1970) and adaptive cluster sampling (Thompson and Seber 1996). Rao’s (1968) theory for sampling from a frame containing an unknown amount of duplication may be regarded as a special case of GWS. Multiple frames can also be handled by GWS and the resulting estimators are simple

but not necessarily efficient compared to the optimal estimators of Hartley (1974) or the PML estimators. The GWS method has wide applicability and deserves the attention of practice.

9. Concluding Remarks

Joe Waksberg's contributions to sample survey theory and methods truly reflect the interplay between theory and practice. Working at the US Census Bureau and later at Westat, he faced real practical problems and produced sound theoretical solutions. For example, his landmark paper (Waksberg 1978) studied an ingenious method (proposed by Warren Mitofsky) for random digit dialing (RDD) that significantly reduces the survey costs compared to dialing numbers completely at random. He presented sound theory to demonstrate its efficiency. The widespread use of RDD surveys is largely due to the theoretical development in Waksberg (1978) and subsequent refinements. Joe Waksberg is one of my heroes in survey sampling and I feel greatly honored to have received the 2005 Waksberg award for survey methodology.

Acknowledgements

I would like to thank David Bellhouse, Wayne Fuller, Jack Gambino, Graham Kalton, Fritz Scheuren and Sharon Lohr for useful comments and suggestions.

References

- Aires, N., and Rosén, B. (2005). On inclusion probabilities and relative estimator bias for Pareto rps sampling. *Journal of Statistical Planning and Inference*, 128, 543-567.
- Andreatta, G., and Kaufmann, G.M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81, 657-666.
- Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.
- Bankier, M.D. (2003). 2001 Canadian Census weighting: switch from projection GREG to pseudo-optimal regression estimation. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Technical Report no. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.
- Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Methodology Branch, Working Paper, Census Operations Section, Social Survey Methods Division, Statistics Canada, Ottawa.
- Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference* (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.
- Bellhouse, D.R., and Rao, J.N.K. (2002). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, 102, 47-58.
- Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.
- Bellhouse, D.R., and Stafford, J.E. (2001). Local polynomial regression in complex surveys. *Survey Methodology*, 27, 197-203.
- Bellhouse, D.R., Chipman, H.A. and Stafford, J.E. (2004). Additive models for survey data via penalized least squares. Technical Report.
- Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Binder, D.A., and Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.
- Binder, D.A., Kovacevic, M. and Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 6-62.
- Brackstone, G. (2002). Strategies and approaches for small area statistics. *Survey Methodology*, 28, 117-123.
- Brackstone, G., and Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 42, 97-114.
- Brewer, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- Brewer, K.R.W., and Hanif, M. (1983). *Sampling With Unequal Probabilities*. New York: Springer-Verlag.
- Buskirk, T.D., and Lohr, S.L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.
- Casady, R.J., and Valliant, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.
- Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- Chambers, R.L., and Skinner, C.J. (Eds.) (2003). *Analysis of Survey Data*. Chichester: Wiley.
- Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 12, 1223-1239.
- Chen, J., Chen, S.Y. and Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53-68.
- Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.
- Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- Cochran, W.G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science*, 30, 262-275.

- Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 191-212.
- Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples from a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- Cochran, W.G. (1953). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Dalenius, T. (1957). *Sampling in Sweden*. Stockholm: Almqvist and Wicksell.
- Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- Deming, W.E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.
- Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.
- Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *The Annals of Mathematical Statistics*, 11, 427-444.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17-26.
- Deville, J., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Durbin, J. (1968). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute*, 36, No. 3, 113-119.
- Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.
- Ernst, L.R. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh), New York: John Wiley & Sons, Inc., 135-169.
- Ernst, L.R. (1999). The maximization and minimization of sample overlap problem: A half century of results. *Bulletin of the International Statistical Institute*, Vol. LVII, Book 2, 293-296.
- Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Fellegi, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS sampling without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington DC, 434-442.
- Fellegi, I.P. (1981). Should the census counts be adjusted for allocation purposes? – Equity considerations. In *Current Topics in Survey Sampling* (Eds. D. Krewski, R. Platek and J.N.K. Rao). New York: Academic Press, 47-76.
- Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, series C, 37, 117-132.
- Fuller, W.A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63, 121-147.
- Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331-345.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Fuller, W.A., and Burneister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.
- Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.
- Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27, 65-74.
- Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, series B*, 17, 269-278.
- Godambe, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, series B*, 28, 310-328.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54, 127-138.
- Graubard, B.I., and Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.
- Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.
- Hájék, J. (1971). Comments on a paper by Basu, D. In *Foundations of Statistical Inference* (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston,
- Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- Hansen, M.H., Dalenius, T. and Tepping, B.J. (1985). The development of sample surveys of finite populations. Chapter 13 in *A Celebration of Statistics*. The ISI Centenary Volume, Berlin: Springer-Verlag.
- Hansen, M.H., Hurwitz, W.N. and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.
- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. I and II. New York: John Wiley & Sons, Inc.
- Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- Hansen, M.H., Hurwitz, W.N., Marks, E.S. and Mauldin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- Hansen, M.H., Hurwitz, W.N., Nisselson, H. and Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.
- Hartley, H.O. (1959). Analytical studies of survey data. In *Volume in Honour of Corrado Gini*, Istituto di Statistica, Rome, 1-32.

- Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36, 99-118.
- Hartley, H.O., and Biemer, P. (1978). The estimation of nonsampling variances in current surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 257-262.
- Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.
- Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- Hartley, H.O., and Rao, J.N.K. (1978). The estimation of nonsampling variance components in sample surveys. In *Survey Measurement* (Ed. N.K. Namboodiri), New York: Academic Press, 35-43.
- Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.
- Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 11-21.
- Holt, D., and Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 300-305.
- Hubback, J.A. (1927). Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (represented in *Sankhyā*, 1946, vol. 7, 281-294).
- Hussain, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.
- Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.
- Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, 129-154.
- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.
- Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, A13, 1919-1939.
- Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changing in the probabilities. *Journal of the American Statistical Association*, 46, 105-109.
- Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.
- Kim, J., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.
- Kim, J.K., Brick, J.M., Fuller, W.A. and Kalton, G. (2004). On the bias of the multiple imputation variance estimator in survey sampling. Technical Report.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L. (1995). The hundred year's wars of survey sampling. *Statistics in Transition*, 2, 813-830.
- Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, series B*, 36, 1-37.
- Kish, L., and Scott, A.J. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389.
- Kott, P.S. (2005). Randomized-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263-277.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- Kruskal, W.H., and Mosteller, F. (1980). Representative sampling IV: The history of the concept in Statistics, 1895-1939. *International Statistical Review*, 48, 169-195.
- Laplace, P.S. (1820). A philosophical essay on probabilities. English translation, Dover, 1951.
- Lavallée, P. (2002). *Le Sondage indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Belgique, Éditions Ellipse, France.
- Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.
- Lehtonen, R., and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.
- Lindley, D.V. (1996). Letter to the editor. *American Statistician*, 50, 197.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury.
- Lohr, S.L., and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 2710280.
- Lohr, S.L., and Rao, J.N.K. (2005). Multiple frame surveys: point estimation and inference. *Journal of the American Statistical Association* (under revision).
- Lu, W.W., Brick, M. and Sitter, R.R. (2004). Algorithms for constructing combined strata grouped jackknife and balanced repeated replication with domains. Technical Report, Westat, Rockville, Maryland.
- Mach, L., Reiss, P.T. and Schiopu-Kratina, I. (2005). The use of the transportation problem in co-ordinating the selection of samples for business surveys. Technical Report HSMD-2005-006E, Statistics Canada, Ottawa.
- Madow, W.G., and Madow, L.L. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.
- Mahalanobis, P.C. (1944). On large scale sample surveys. *Philosophical Transactions of the Royal Society*, London, Series B, 231, 329-451.
- Mahalanobis, P.C. (1946a). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.
- Mahalanobis, P.C. (1946b). Sample surveys of crop yields in India. *Sankhyā*, 7, 269-280.

- McCarthy, P.J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37, 239-264.
- Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.
- Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.
- Murthy, M.N. (1964). On Mahalanobis' contributions to the development of sample survey theory and methods. In *Contributions to Statistics: Presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday*, Calcutta, Statistical Publishing Society: 283-316.
- Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.
- Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- Ohlsson, E. (1995). Coordination of samples using permanent random members. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), New York: John Wiley & Sons, Inc., 153-169.
- O'Muircheartaigh, C.A., and Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: A review. *Bulletin of the International Statistical Institute*, Invited Papers, 49, No. 1, 465-493.
- Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- Owen, A.B. (2002). *Empirical Likelihood*. New York: Chapman & Hall/CRC.
- Park, M., and Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology*, 31, 85-93.
- Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data* (Eds. R.L. Chambers and C.J. Skinner), Chichester: Wiley, 175-195.
- Raj, D. (1956). On the method of overlapping maps in sample surveys. *Sankhyā*, 17, 89-98.
- Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28, 47-60.
- Rao, J.N.K. (1968). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.
- Rao, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Proceedings of the workshop on uses of auxiliary information in surveys*, Statistics Sweden.
- Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- Rao, J.N.K. (1996). Developments in sample survey theory: An appraisal. *The Canadian Journal of Statistics*, 25, 1-21.
- Rao, J.N.K. (2000). Variance estimation in the presence of imputation for missing data. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 599-608.
- Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken: Wiley.
- Rao, J.N.K. (2004). Empirical likelihood methods for sample survey data: An overview. *Proceedings of the Survey Methods Section, SSC Annual Meeting*, in press.
- Rao, J.N.K. (2005). Re-sampling variance estimation with imputed survey data: overview. *Bulletin of the International Statistical Institute*.
- Rao, J.N.K., and Bellhouse, D.R. (1990). History and development of the theoretical foundations of survey based estimation and analysis. *Survey Methodology*, 16, 3-29.
- Rao, J.N.K., and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.
- Rao, J.N.K., and Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-64.
- Rao, J.N.K., and Singh, M.P. (1973). On the choice of estimators in survey sampling. *Australian Journal of Statistics*, 15, 95-104.
- Rao, J.N.K., and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics – Theory and Methods*, 33, 2087-2095.
- Rao, J.N.K., and Wu, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: Some recent work. *Bulletin of the International Statistical Institute*.
- Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-491.
- Rao, J.N.K., Jocelyn, W. and Hidiroglou, M.A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19.
- Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling. *Survey Methodology*, 29, 107-128.

- Rao, J.N.K., Yung, W. and Hidirolou, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā, Series A*, 64, 364-378.
- Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-375.
- Rivest, L.-P. (2002). A generalization of the Lavallée and Hidirolou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.
- Roberts, G., Rao, J.N.K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- Rosén, B. (1991). Variance estimation for systematic pps-sampling. Technical Report, Statistics Sweden.
- Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.
- Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- Royall, R.M., and Cumberland, W.G. (1981). An empirical study of the ratio estimate and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- Royall, R.M., and Herson, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68, 880-889 and 890-893.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- Rubin-Bleuer, S., and Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, (to appear).
- Salehi, M., and Seber, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacements. *Biometrika*, 84, 209-219.
- Särndal, C.-E. (1996). Efficient estimators with variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- Särndal, C.-E., Swenson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- Särndal, C.-E., Swenson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Schabenberger, O., and Gregoire, T.G. (1994). Competitors to genuine rps sampling designs. *Survey Methodology*, 20, 185-192.
- Schaible, W.L. (Ed.) (1996). *Indirect Estimation in U.S. Federal Programs*. New York: Springer
- Scott, A., and Davis, P. (2001). Estimating interviewer effects for survey responses. *Proceedings of Statistics Canada Symposium 2001*.
- Shao, J. (2002). Resampling methods for variance estimation in complex surveys with a complex design. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc., 303-314.
- Shao, J., and Tu, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer Verlag.
- Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 27, 33-44.
- Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.
- Singh, A.C., and Wu, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 69-77.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-14.
- Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence interval for quantiles. *Statistics & Probability Letters*, 55, 353-358.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.
- Smith, T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society, Series A*, 139, 183-204.
- Smith, T.M.F. (1994). Sample surveys 1975-1990; an age of reconciliation? *International Statistical Review*, 62, 5-34.
- Stehman, S.V., and Overton, W.S. (1994). Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.
- Sukhatme, P.V. (1947). The problem of plot size in large-scale yield surveys. *Journal of the American Statistical Association*, 42, 297-310.
- Sukhatme, P.V. (1954). *Sampling Theory of Surveys, with Applications*. Ames: Iowa State College Press.
- Sukhatme, P.V., and Panse, V.G. (1951). Crop surveys in India – II. *Journal of the Indian Society of Agricultural Statistics*, 3, 97-168.
- Sukhatme, P.V., and Seth, G.R. (1952). Non-sampling errors in surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.
- Thomas, D.R., and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.
- Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.
- Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *International Statistical Review*, 66, 303-322.
- Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461-493, 646-683.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the U.S. Census Bureau. *Journal of Official Statistics*, 14, 119-135.

- Wang, N., and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.
- Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.
- Wu, C., and Rao, J.N.K. (2004). Empirical likelihood ratio confidence intervals for complex surveys. Submitted for publication.
- Wu, C., and Rao, J.N.K. (2005). Empirical likelihood approach to calibration using survey data. Paper presented at the 2005 International Statistical Institute meetings, Sydney, Australia.
- Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Griffin.
- Zarkovic, S.S. (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society, Series A*, 119, 336-338.

Hot Deck Imputation for the Response Model

Wayne A. Fuller and Jae Kwang Kim¹

Abstract

Hot deck imputation is a procedure in which missing items are replaced with values from respondents. A model supporting such procedures is the model in which response probabilities are assumed equal within imputation cells. An efficient version of hot deck imputation is described for the cell response model and a computationally efficient variance estimator is given. An approximation to the fully efficient procedure in which a small number of values are imputed for each nonrespondent is described. Variance estimation procedures are illustrated in a Monte Carlo study.

Key Words: Nonresponse; Fractional imputation; Response probability; Replication variance estimation.

1. Introduction

Imputation is used in sample surveys as a method of handling item nonresponse. In hot deck imputation, the imputed values are functions of the respondents in the current sample. Sande (1983) and Ford (1983) contain descriptions of hot deck imputation. Kalton and Kasprzyk (1986) and Little and Rubin (2002) review various imputation procedures.

In one version of hot deck imputation, the imputed value is the value of a respondent in the same imputation cell, where the imputation cells form an exhaustive and mutually exclusive subdivision of the population. In random hot deck imputation, respondents are assigned values at random from respondents in the same imputation cell. The record providing the value is called the *donor* and the record with the missing value is called the *recipient*.

The variance of the imputed estimator is generally larger than the complete sample variance because nonresponse reduces sample size and because the imputed estimator may contain a component due to random imputation. Rao and Shao (1992) proposed an adjusted jackknife method for hot-deck imputation where the first phase units are selected with-replacement. Rao and Sitter (1995) discussed the adjusted jackknife variance estimation method for ratio imputation. Rao (1996) and Sitter (1997) applied the adjusted jackknife method to regression imputation. Shao, Chen and Chen (1998) apply the idea of Rao and Shao (1992) to the balanced repeated replication method. Shao and Steel (1999) propose variance estimation for survey data with composite imputation, where more than one imputation method is used, and the sampling fractions are included in the variance expressions. Yung and Rao (2000) applied the adjusted jackknife method to imputed estimators constructed with a poststratified sample. Rubin (1987) and

Rubin and Schenker (1986) suggested multiple imputation procedures. Tollefson and Fuller (1992), and Särndal (1992) proposed imputation methods and corresponding variance estimators. Kim and Fuller (2004) studied the use of fractional imputation for the model in which observations in an imputation cell are independently and identically distributed.

In this paper, we consider hot deck imputation for a population divided into imputation cells. The response model is described in section 2. In section 3, we introduce fully efficient fractional imputation and present a variance estimation method for the imputation estimator, under the assumptions that the probability of nonresponse is constant within a cell. In section 4 we suggest a modification of the fully efficient method that uses a smaller number of donors. In section 5, an example is introduced to illustrate the actual implementation of the proposed method. In section 6, results of a simulation study are reported. Summary is presented in the last section.

2. Basic Setup

Consider a population of N elements identified by a set of indices $U = \{1, 2, \dots, N\}$. Associated with each unit i in the population there is a study variable y_i and a vector \mathbf{x}_i of auxiliary information. The set of vectors, (y_i, \mathbf{x}_i) , $i = 1, 2, \dots, N$, is denoted by F .

Let A denote the indices of the elements in a sample selected by a set of probability rules called the *sampling mechanism*. Let the population quantity of interest be θ_N , let $\hat{\theta}$ be a full sample, linear-in- y , estimator of θ_N , and write

$$\hat{\theta} = \sum_{i \in A} w_i y_i. \quad (1)$$

1. Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA, 50011 U.S.A.; Jae Kwang Kim, Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea.

If w_i is the inverse of the selection probability, then $\hat{\theta}$ is unbiased for the population total.

Let A_R and A_M denote the set of indices of the sample respondents and sample nonrespondents, respectively. Define the response indicator function

$$R_i = \begin{cases} 1 & \text{if } i \in A_R \\ 0 & \text{if } i \in A_M \end{cases} \quad (2)$$

and let $\mathbf{R} = \{(i, R_i); i \in A\}$. The distribution of \mathbf{R} is called the *response mechanism*.

Assume that the finite population U is made up of G imputation cells, where the set of elements in cell g is U_g . Let n_g be the number of sample elements in imputation cell g and let $r_g, r_g > 0$, be the number of respondents in imputation cell g . Assume the within-cell uniform response model in which the r_g responses in a cell are equivalent to a Poisson sample selected with equal probabilities from the n_g elements.

Fractional imputation is a procedure in which more than one donor is used per recipient. Kalton and Kish (1984) suggested fractional imputation as an efficient imputation procedure. The method was discussed by Fay (1996). Let d_{ij} be the number of times that y_i is used as donor for the missing y_j and define $\mathbf{d} = \{d_{ij}; i \in A_R, j \in A_M\}$. The distribution of \mathbf{d} is called the *imputation mechanism*. Let w_{ij}^* be the factor applied to the original weight for element j when y_i is used as a donor for element j . For element $j, j \in A_M$,

$$Y_j = \sum_{i \in A_R} w_{ij}^* y_i \quad (3)$$

is the weighted mean of the respondent values. The factor w_{ij}^* is called the *imputation fraction*. It is the fraction that donor i donates for the missing item y_j . Note that $w_{ii}^* = 1$ for $i \in A_R$ and $w_{ij}^* = 0$ for $i \neq j, i, j \in A_R$. The sum of the imputation fractions for a missing item is restricted to equal one,

$$\sum_{i \in A_R} w_{ij}^* = 1, \quad \forall j \in A. \quad (4)$$

An estimator with the imputed values defined in (3) and some $w_{ij}^* < 1$ is called a *fractionally imputed estimator*.

A linear-in- y imputation estimator can be written in the form

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{j \in A} w_j w_{ij}^* \right) y_i \quad (5)$$

$$= \sum_{i \in A_R} \alpha_i y_i, \quad (6)$$

where the notation $A =: B$ means that B is defined to be equal to A . The sum of $w_{ij}^* w_j$ over all recipients for which i is a donor (including acting as a donor for itself), denoted by α_i , is the total weight of donor i . If a responding unit i is not used as a donor, except for itself, then $\alpha_i = w_i$.

3. Fully Efficient Fractional Imputation

Assume all elements in an imputation cell have the same probability of responding and assume the responses are independent. Then the overall distribution of an imputed estimator under the response model can be obtained by using the probability structure of multiple phase sampling, where the response model is treated as the second phase sampling mechanism.

If the response probabilities in a cell are uniform, then a reasonable estimator of the total is the weighted sum of ratio estimators

$$\hat{\theta}_{FE} = \sum_{g=1}^G \left(\sum_{i \in A \cap U_g} w_i \right) \frac{\sum_{i \in A_R \cap U_g} w_i y_i}{\sum_{i \in A_R \cap U_g} w_i}. \quad (7)$$

In the context of two phase sampling, Kott and Stukel (1997) call the estimator (7) a reweighted expansion estimator. The estimator (7) is called fully efficient because it contains no variability due to random selection of donors. If the w_i are the same for all elements in a cell, the ratio

$$\left(\sum_{i \in A_R \cap U_g} w_i \right)^{-1} \sum_{i \in A_R \cap U_g} w_i y_i \quad (8)$$

is a simple mean and, hence, unbiased for the cell mean given that there is at least one respondent in the cell. If the w_i in a cell are not equal, then (8) is subject to ratio bias. It is possible for the number of elements in a cell, n_g , to be positive and the number of respondents, r_g , to be zero. If this occurs in practice, cells will be collapsed.

The large sample properties of the estimator can be obtained for a sequence of populations and samples. Assume the population is composed of G_v mutually exclusive and exhausted cells, where v is the index of the sequence. Assume the variance of a full sample estimator of the mean is $O(n_v^{-1})$, where n_v is the size of the sample selected from the v^{th} population. Assume responses are independent. Then, under regularity conditions, the procedures used by Kim, Navarro and Fuller (2005) in the proof of their Theorem 2.1 can be used to show that estimator (7) satisfies

$$\hat{\theta}_{FE_v} = \hat{\theta}_v + \sum_{g_v=1}^{G_v} \sum_{i \in A_{g_v}} w_{i_v} (\pi_{g_v}^{-1} R_{i_v} - 1) e_{i_v} + o_p(n_v^{-1/2} N_v), \quad (9)$$

where $e_{iv} = y_{iv} - \bar{Y}_{gv}$, A_{gv} is the set of sample indices in the g^{th} cell for the v^{th} sample, \bar{Y}_{gv} is the population mean of the y -variable in cell gv of population F_v , π_{gv} is the probability that an element in cell gv responds, and F_v denotes the v^{th} population. Also

$$V(\tilde{\theta}_{\text{FEV}} | F_v) = V(\hat{\theta}_v | F_v) + E \left\{ \sum_{g_v=1}^{G_v} \pi_{gv}^{-1} (1 - \pi_{gv}) \sum_{i \in A_{gv}} w_{iv}^2 e_{iv}^2 | F_v \right\}, \quad (10)$$

where

$$\tilde{\theta}_{\text{FEV}} = \hat{\theta}_v + \sum_{g_v=1}^{G_v} \sum_{i \in A_{gv}} w_{iv} (\pi_{gv}^{-1} R_{iv} - 1) e_{iv}.$$

The estimator (7) can be implemented by using fractional imputation in which every responding unit in an imputation cell is used as a donor for every nonrespondent in the cell. Then, the estimator (7) can be written as the fractionally imputed estimator

$$\hat{\theta}_{\text{FEFI}} = \sum_{g=1}^G \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i, \quad (11)$$

where $w_j w_{ij}^*$ is the weight of donor i for recipient j , w_{ij}^* is the imputation fraction of donor i for recipient j defined in (3), and

$$w_{ij}^* = \begin{cases} \left(\sum_{s \in A_R \cap U_g} w_s \right)^{-1} w_i R_i & \text{if } R_j = 0 \\ 1 & \text{if } R_j = 1 \text{ and } i = j. \end{cases} \quad (12)$$

The estimator (11) with w_{ij}^* of (12), algebraically equivalent to (7), is called the *fully efficient fractionally imputed* (FEFI) estimator. The fractionally imputed estimator has the advantage that functions of y such as the fraction less than a given number can be directly estimated from the fractionally imputed data set.

To consider replication variance estimation, let a replication variance estimator for the complete sample be

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^L c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \quad (13)$$

where $\hat{\theta}^{(k)}$ is the k^{th} estimate of θ_N based on the observations included in the k^{th} replicate, L is the number of replicates, and c_k is a factor associated with replicate k determined by the replication method. For a discussion of replication for survey samples see Krewski and Rao (1981) and Rao, Wu and Yue (1992). When the original estimator $\hat{\theta}$ is a linear estimator of the form (1), the k^{th} replicate estimate of $\hat{\theta}$ can be written

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i, \quad (14)$$

where $w_i^{(k)}$ denotes the replicate weight for the i^{th} unit of the k^{th} replication.

A proposed replicate for the estimator $\hat{\theta}_{\text{FEFI}}$ is

$$\begin{aligned} \hat{\theta}_{\text{FEFI}}^{(k)} &= \sum_{g=1}^G \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} \\ &= \sum_{g=1}^G \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i. \end{aligned} \quad (15)$$

Using the replicates (15), the replicate variance estimator can be written as

$$\hat{V}_{\text{FEFI}} = \sum_{k=1}^L c_k (\hat{\theta}_{\text{FEFI}}^{(k)} - \hat{\theta}_{\text{FEFI}})^2. \quad (16)$$

The replicates in (15) can be computed in two steps. First, create the usual replicate by defining the weights $w_i^{(k)}$ for every element. Second, for a nonrespondent, the replicate imputation fraction for donor i to recipient j is

$$w_{ij}^{*(k)} = \frac{w_i^{(k)}}{\sum_{s \in A_R \cap U_g} w_s^{(k)}}.$$

Note that the sum of the fractional replication weights of the donor records for each recipient is the same as the replication weight for that unit in a complete sample.

The suggested procedure is closely related to the Rao and Shao (1992) variance estimator. See also Yung and Rao (2000). However, the use of fractional imputation greatly simplifies variance estimation. In the creation of replicates, only the weights on the imputed values are changed. No recomputing of imputed values is required, and once computed, the replicate weights can be used for any smooth function of the vector y . Also, the fractional replicates make the estimator (16) appropriate for a vector of y -variables.

Theorem 3.1 of Kim, Navarro and Fuller (2005) can be used to show that, given a consistent full sample replication procedure,

$$\begin{aligned} \hat{V}_{\text{FEFI}} &= V(\tilde{\theta}_{\text{FEV}} | F_v) \\ &- N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1} (1 - \pi_{gv}) e_{iv}^2 + o_p(n_v^{-1}), \end{aligned} \quad (17)$$

where $\tilde{\theta}_{\text{FEV}}$ is defined in (10), and the distribution is with respect to the sampling and response mechanisms.

If the finite population correction can be ignored, the estimator (16) is consistent for $V\{\hat{\theta}_{\text{FE}}\}$. If the sample size is large relative to N , then an estimator of

$$N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1} (1 - \pi_{gv}) e_{iv}^2$$

should be added to (16).

The imputation and variance estimation procedure outlined for the response model also produces consistent estimators for the cell mean model. Under the cell mean model, the elements within a cell of the finite population are a realization of independently and identically distributed random variables. The imputation procedure based on the response model is not necessarily fully efficient for the population mean under the cell mean model, but it can be shown that the estimator of the mean and the estimator of the variance of the estimated mean are consistent.

4. Approximations to the Fully Efficient Procedure

In the previous sections, the estimator $\hat{\theta}_{\text{FEFI}}$ was constructed to produce zero imputation variance. The implementation of the fractional imputation procedure, as described in (11), could require the use of a large number of donors for each recipient. Therefore, we outline a procedure with a fixed number of donors per recipient that is fully efficient for the grand total, but not necessarily fully efficient for subpopulations. The procedure assigns donors to produce small between-recipient variance of imputed values and modifies the weights of donors to attain full efficiency for the total.

Suppose that M donors are to be assigned to each recipient. We suggest donors be assigned to recipients to approximate the distribution of all respondents in the cell. One possible selection method is to select a stratified sample for each recipient. A second possibility is to use systematic sampling with probability proportional to the weights to select donors for each recipient. Initial fractions w_{ij0}^* are assigned to the donated values. For systematic sampling with equal weights, the initial w_{ij0}^* is M^{-1} .

After the donors are assigned, the initial fractions, w_{ij0}^* are adjusted so that the sum of the weights gives the fully efficient estimator of the mean of y , and such that the estimated cumulative distribution function based on the weights approximates the fully efficient estimator of the cumulative distribution function. The modification of weights using regression has been suggested by Fuller (1984, 2003). Chen, Rao and Sitter (2000) discussed an efficient imputation method that changes the imputed values rather than the weights. Let $\mathbf{z}_{g \cdot j} = (z_{g \cdot j1}, z_{g \cdot j2}, \dots, z_{g \cdot j\alpha})$ be a vector defined by

$$\begin{aligned} z_{g \cdot j1} &= y_j \\ z_{g \cdot j2} &= 1 \quad \text{if } y_j \leq L_2 \\ &= 0 \quad \text{otherwise} \\ &\vdots \\ z_{g \cdot j\alpha} &= 1 \quad \text{if } L_{\alpha-1} < y_j \leq L_\alpha \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

where $L_2, L_3, \dots, L_\alpha$ divide the range of observed y in cell g into $\alpha-1$ sections. The number of sections that can be used depends on the numbers and type of observations in the cell, the number of recipients and the number of donors per recipient. If the number of donors per recipient is large, it is possible to adjust the set of weights for each recipient so that the sum of w_{ij}^* over i is one for every j and the sum of $w_{ij}^* y_i$ over i is the fully efficient estimator for every j . In most cases the weights will be adjusted so that the sum of the w_{ij}^* over i is one for every j and the cell means of the imputed values are equal to the fully efficient estimator.

Let $\bar{\mathbf{z}}_{\text{FE},g}$ denote the fully efficient estimator for cell g . Using regression procedures, the modified w_{ij}^* , modified to give the fully efficient cell mean of \mathbf{z} , are

$$w_{ij}^* = w_{ij0}^* + (\bar{\mathbf{z}}_{\text{FE},g} - \bar{\mathbf{z}}_g^*) \mathbf{S}_{\mathbf{z}\mathbf{z}}^{-1} w_{ij0}^* (\bar{\mathbf{z}}_{g[i]j} - \bar{\mathbf{z}}_{g \cdot j})', \quad (18)$$

where

$$\begin{aligned} \mathbf{S}_{\mathbf{z}\mathbf{z}} &= \sum_{j \in A_{Lg}} b_j \sum_{i \in A_{Rg}} w_{ij0}^* (\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g \cdot j})' (\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g \cdot j}) d_{ij}, \\ \bar{\mathbf{z}}_{g \cdot j} &= \sum_{i \in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j} d_{ij}, \\ \bar{\mathbf{z}}_g^* &= \sum_{j \in A_{Lg}} b_j \sum_{i \in A_{Rg}} w_{ij0}^* \mathbf{z}_{g[i]j} d_{ij}, \\ b_j &= \left(\sum_{s \in A_{Lg}} w_s \right)^{-1} w_j, \end{aligned}$$

A_{Lg} is the set of indexes of recipients in cell g , $\mathbf{z}_{g[i]j} = \mathbf{z}_{gi}$ is the value imputed from donor i for recipient j , and $\bar{\mathbf{z}}_{g \cdot j}$ is the weighted mean of the imputed values for recipient j using the initial w_{ij0}^* .

To estimate the variance, replicates are created so that the weights on the donors reflect the effect of the deletion of an element on the fully efficient estimator. We use the words "deletion" and "delete" to identify the element chosen for principal weight modification for replication variance estimation.

Let $w_i^{(k)}$ be the weight assigned to element i for the k^{th} replicate for variance estimation of the full sample estimator. Then the replicate for the fully efficient mean of y for cell g is

$$\bar{z}_g^{(k)} = \left[\sum_{i \in A_{Rg}} w_i^{(k)} \right]^{-1} \sum_{i \in A_{Rg}} w_i^{(k)} z_i. \tag{19}$$

Replicate fractions are assigned to donors in cell g so that the replicate estimate of the cell mean is $\bar{z}_g^{(k)}$. Initial fractional weights $w_{ij0}^{*(k)}$ are assigned where $w_{ij0}^{*(k)}$ is small, but positive, if i is a deleted unit for replicate k . The final fractional weights $w_{ij}^{*(k)}$ are computed using the procedure of (18) with $\bar{z}_g^{(k)}$ replacing $\bar{z}_{FE,g}$ and $w_{ij0}^{*(k)}$ replacing w_{ij0}^* . The procedure simulates the effect of deleting a single element on the fully efficient estimator.

5. An Artificial Example

In this section, we present an example with artificial data to illustrate the implementation of the proposed method. Suppose that two study variables, x and y , are observed in a sample of size $n = 10$ obtained by simple random sampling. Variable x is a categorical variable with three categories, say 1, 2, and 3, and variable y is a continuous variable. Both variables have item nonresponse and there is a set of imputation cells for each variable. Table 5.1 shows the sample observations, where nonresponse is denoted by M in the table. We use a weight of one to simplify the presentation. Divide by ten to obtain weights for the mean.

Table 5.1
An Illustrative Data Set

Observation	Weight	Cell for x	Cell for y	x	y
1	1	1	1	1	7
2	1	1	1	2	M
3	1	1	2	3	M
4	1	1	1	M	14
5	1	1	2	1	3
6	1	2	1	2	15
7	1	2	2	3	8
8	1	2	1	3	9
9	1	2	2	2	2
10	1	2	1	M	M

Because the x variable is a categorical variable with three categories, using three fractions for fractional imputation gives fully efficient estimators for the distribution of the x -variable. Thus the weights in Table 5.2 for the three imputed values of x for observation four are the fractions for the three categories in x -cell one.

If a subset of donors is to be used for each recipient, a controlled method of selecting donors, such as systematic sampling, is suggested. In our simple illustration we could easily use fractional imputation with all four y responses in cell 1, but to illustrate the regression adjustment we use only three. See Table 5.2.

Several approaches are possible for the situation in which two items are missing, including the definition of a third set

of imputation cells for such cases. Because of the small size of our illustration, we impute under the assumption that x and y are independent within cells. Thus we impute four values for observation ten. For each of the two possible values of x we impute two possible values for y . One of the pair of imputed y -values is chosen to be less than the mean of responses and one is chosen to be greater than the mean. See the imputed values for observation 10 in Table 5.2.

Table 5.2
Fractional Weights for Means

Observation	Weight	Donor for y	Cell for x	Cell for y	x	y
1	1.0000		1	1	1	7
2	0.2886	1	1	1	2	7
2	0.3960	6	1	1	2	15
2	0.3154	8	1	1	2	9
3	0.3333	5	1	2	3	3
3	0.3333	7	1	2	3	8
3	0.3334	9	1	2	3	2
4	0.5000	1	1	1	1	14
4	0.2500		1	1	2	14
4	0.2500		1	1	3	14
5	1.0000	1	2	1	1	3
6	1.0000		2	1	2	15
7	1.0000		2	2	3	8
8	1.0000		2	1	3	9
9	1.0000		2	2	2	2
10	0.2247	8	2	1	2	9
10	0.2753	4	2	1	2	14
10	0.2095	1	2	1	3	7
10	0.2905	6	2	1	3	15

Initial fractions of one third are assigned to the three imputed values for observations three and four, and initial fractions of one fourth are assigned to the four imputed values for observation ten. The fractional weights are then adjusted using the regression method of equation (18) to give the FEFI mean of y as the estimator, where the fully efficient estimator for the mean of y is

$$\bar{y}_{FE} = \sum_{g=1}^2 \frac{n_g}{n} \bar{y}_{Rg} = 8.4833.$$

We restrict the weights for observation 10 so that the estimated fractions for the two categories of x are the cell fractions. Then, because the weighted mean for the categorical variable is controlled for each individual, the vector z contains only the y -variable. Table 5.2 gives the final fractional weights computed with the regression weighting.

An analyst can use the data set of Table 5.2 and any full-sample computer program to compute estimates of functions of y and x , such as the mean of y for the x categories. The fractional data set is fully efficient for any function of the x -variable and is also fully efficient for the mean of the y -variable.

For jackknife variance estimation, we repeat the weight calculation for each replicate. The replicate estimates of the cell means of y are given in Table 5.3 and the replicate

estimates of the category fractions for x are given in Table 5.4. The values in Table 5.3 and in Table 5.4 are used as the control totals $\bar{z}_{FE,g}$ in the regression weighting. We used $w_{ij0}^{*(k)} = 3^{-1}$ as the initial value of the replication fractions for observation two and $w_{ij0}^{*(k)} = 4^{-1}$ for observation ten.

Table 5.5 contains the jackknife weights for the fractionally imputed data set of Table 5.2. The replicate weights are used in the same way as replicates for a full sample. They are appropriate, with the caveats of the next section, for any statistic for which the full sample jackknife is appropriate. Thus the procedure is particularly appealing for a general purpose data set, because no additional computations are required of the analyst.

The fully efficient estimator of the mean of y is obtained by treating the respondents as the second phase of a two phase sample. A two-phase variance estimator is

$$\hat{V} = \frac{1}{n} \sum_{g=1}^2 \frac{n_g}{n} (\bar{y}_{Rg} - \bar{y}_{FE})^2 + \sum_{g=1}^2 \left(\frac{n_g}{n} \right)^2 \frac{1}{r_g} s_{Rg}^2 = 3.043,$$

where s_{Rg}^2 is the within cell sample variance for cell g . If we use the replication weights in Table 5.5, the replication variance estimate for the mean of y is

$$\hat{V}_{JK}(\bar{y}_{FI}) = \sum_{k=1}^{10} 0.9 (\bar{y}_{FI}^{(k)} - \bar{y}_{FI})^2 = 3.078.$$

The difference between the linearized variance estimator and the jackknife variance estimator is

$$\sum_{g=1}^2 \left(\frac{r_g}{r_g - 1} \frac{n-1}{n} - 1 \right) s_{Rg}^2.$$

Thus, the jackknife variance estimator slightly overestimates the true variance in this example.

Table 5.3
Jackknife Replicates of Cell Mean of y -variable

Cell	Replicate									
	1	2	3	4	5	6	7	8	9	10
1	12.67	11.25	11.25	10.33	11.25	10.00	11.25	12.00	11.25	11.25
2	4.33	4.33	4.33	4.33	5.00	4.33	2.50	4.33	5.50	4.33

Table 5.4
Jackknife Replicates of Cell Mean of the Dummy Variables of x -variable

Cell	Level of x	Replicate									
		1	2	3	4	5	6	7	8	9	10
1	1	0.33	0.67	0.67	0.50	0.33	0.50	0.50	0.50	0.50	0.50
	2	0.33	0.00	0.33	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	3	0.33	0.33	0.00	0.25	0.33	0.25	0.25	0.25	0.25	0.25
2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.50	0.50	0.50	0.50	0.50	0.33	0.67	0.67	0.33	0.50
	3	0.50	0.50	0.50	0.50	0.50	0.67	0.33	0.33	0.67	0.50

Table 5.5
Jackknife Weights for Fractional Imputation

Obs.	Replicate									
	1	2	3	4	5	6	7	8	9	10
1	0	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111
2	0.1664	0	0.3206	0.4205	0.3206	0.4563	0.3206	0.2392	0.3206	0.2724
2	0.6559	0	0.4400	0.3002	0.4400	0.2500	0.4400	0.5540	0.4400	0.5075
2	0.2888	0	0.3505	0.3904	0.3505	0.4048	0.3505	0.3179	0.3505	0.3312
3	0.3706	0.3706	0	0.3706	0.3226	0.3706	0.5018	0.3706	0.2867	0.3706
3	0.3697	0.3697	0	0.3697	0.5018	0.3697	0.0090	0.3697	0.6004	0.3697
3	0.3708	0.3708	0	0.3708	0.2867	0.3708	0.6003	0.3708	0.2240	0.3708
4	0.3703	0.7407	0.7407	0	0.3703	0.5556	0.5556	0.5556	0.5556	0.5556
4	0.3704	0	0.3704	0	0.3704	0.2777	0.2777	0.2777	0.2777	0.2777
4	0.3704	0.3704	0	0	0.3704	0.2778	0.2778	0.2778	0.2778	0.2778
5	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111	1.1111	1.1111
6	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111	1.1111
7	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111	1.1111
8	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111	1.1111
9	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	1.1111	0	1.1111
10	0.1624	0.2777	0.2777	0.3061	0.2777	0.2286	0.3474	0.3013	0.1520	0
10	0.3931	0.2778	0.2778	0.2494	0.2778	0.1417	0.3934	0.4395	0.2185	0
10	0.0932	0.2778	0.2778	0.3231	0.2778	0.4400	0.1483	0.0746	0.3171	0
10	0.4623	0.2778	0.2778	0.2324	0.2778	0.3008	0.2220	0.2957	0.4235	0

6. Simulation Studies

6.1 Study Parameters

To study the properties of the imputation procedure we conducted a Monte Carlo study. The sample is a stratified sample with two elements per stratum and two imputation cells, where the cells cut across the strata. Cell one is 20% of the population in strata 1–25 and 80% of the population in strata 26–50. The probability of response is 0.7 for cell one and 0.5 for cell two. Two variables are considered. The variable D is always observed and defines a subpopulation. The probability that $D = 1$ is 0.25 for cell one and 0.40 for cell two. The variable y is subject to nonresponse with constant within-cell response probabilities. The variable D is independent of y and of the response probability. The variable y is normally distributed, where the parameters for a population of 50 strata are given in Table 5.1. In the data generating model of Table 6.1, there are no stratum effects. The parameters of interest are: θ_1 = mean of y , θ_2 = mean of y for $D=1$, θ_3 = fraction of Y 's less than two, θ_4 = fraction of Y 's less than one.

Table 6.1
Parameter Set A

		Cell One		Cell Two	
Strata	Element Weight	Mean	Variance	Mean	Variance
1–25	0.01	0.4	0.36	1.6	0.36
26–50	0.01	0.4	0.36	1.6	0.36

6.2 Estimation Procedures

In the simulation $M = 5$ and $M = 3$ donors were used per recipient. Systematic samples were selected to serve as donors for each recipient. If the number of respondents in the cell is less than M , every respondent was used as a donor for every recipient and the w_{ij}^* are proportional to the original w_i of the respondents. If there are more than M respondents in a cell, the donors are ordered by size and numbered from one to r_g . Then the donors are placed in the order 1, 3, 5, ..., r_g , r_{g-1} , r_{g-3} , ..., 2 for r_g odd and the order 1, 3, 5, ..., r_{g-1} , r_g , r_{g-2} , ..., 2 for r_g even. The cumulated sums of the weights are formed and m_g systematic samples of size M are selected, where $m_g = n_g - r_g$. The cumulative sums are normalized so that the grand sum is one, a random number, R_{Ng} , between zero and $0.2m_g$ is selected and the m_g samples are the systematic samples of size M defined by the donor associated with $R_{Ng} + 0.2(s-1) + (t-1)m_g^{-1}$, $s=1, 2, 3, 4, 5$ for recipients $t=1, 2, \dots, m_g$. The initial imputation fraction for each donor is $w_{ij}^* = M^{-1}$.

The initial imputation fractions are modified using the regression procedure of (18). The donors in a cell were ordered from smallest to largest and the cumulative sum of the weights formed. Let

$$S_{g,wt} = \sum_{i=1}^I w_{[i]}, i \in A_{Rg}, \quad (20)$$

where $w_{[i]}, i=1, 2, \dots, r_g$, is the weight of $y_{g,(i)}$ and $y_{g,(1)} \leq \dots \leq y_{g,(n)}$ are the ordered y -values in cell g . To define the boundaries of groups to be used to create indicator functions, let t_{*s} be the t for which

$$\max \{S_{g,wt} : S_{g,wt} \leq 0.2sS_{gw}\}$$

for $s=1, 2, 3, 4$, where S_{gw} is the total of the weights of the donors in cell g . Define

$$\begin{aligned} z_{gi,s+1} &= 1 \quad \text{if } y_i \leq y_{g,(t_{*s})} \text{ and } i \in A_{Rg} \\ &= 0 \quad \text{otherwise} \end{aligned} \quad (21)$$

for $s=1, 2, 3, 4$ and let $\mathbf{z}_{gj} = (y_{gj1}, z_{gj2}, \dots, z_{gj5})$. The regression modified imputed estimator of the mean for each of the five variables in the \mathbf{z} -vector is the fully efficient estimator of the respective mean.

The k -deleted FE estimator of the cell mean of \mathbf{z} is defined in (19). The initial fractional weight for donor k to element j is set at $w_{kj0}^{*(k)} = 0.01w_{kj}$. This initial weight assures that the final weight will be small, but permits regression adjustment. The final $w_{ij}^{*(k)}$ are computed using the regression procedure of (18) using the initial weight $w_{ij0}^{*(k)}$.

6.3 Monte Carlo Results

The Monte Carlo results for 5,000 samples generated by the parameters of Table 6.1 are given in Table 6.2 and Table 6.3. Results are given for the full sample, for fractional imputation with 5 donors, fractional imputation with three donors, and for multiple imputation (MI) using the Approximate Bayesian Bootstrap (ABB) of Rubin and Schenker (1986) with $M=5$ and ABB with $M=3$. Both the FI and MI procedures are unbiased for all four parameters of Table 6.2. The last column of Table 6.2 gives the Monte Carlo variance of the estimator divided by the Monte Carlo variance of the FI procedure with $M=5$, expressed in percent. The FI procedure is five to ten percent more efficient than MI with $M=5$ and 9 to 13 percent more efficient than MI with $M=3$.

Under the model, the mean of the observed values is not the best estimator of the domain mean. In this example, the FI estimator is about as efficient as the full sample estimator. The effect of a smaller number of observations is balanced by the use of a superior estimator of the mean for the domain. Under the model, the domain indicator is independent of the y values, given the cell. Therefore it is efficient to use all values in the cell as donors, not just respondents in the domain.

The properties of the variance estimators are given in Table 6.3. The column headed "Relative Mean" gives the Monte Carlo estimated mean of the estimated variances

divided by the Monte Carlo estimated variance, where the Monte Carlo estimated variance is given in Table 6.2. Both variance estimation procedures appear to be nearly unbiased for the variance of the mean. The relative variance of the MI variance estimator for $M = 5$ is nearly twice that of the FI variance estimator for $M = 5$. For $M = 3$, the MI variance estimator is more than three times that for FI. The MI variance estimator has a large variance because the variance due to missing observations is estimated with four degrees-of-freedom for $M = 5$ and with two-degrees-of freedom for $M = 3$.

The MI variance estimator for the domain mean is seriously biased. This property was first identified by Fay

(1991, 1992) and studied by Meng (1994) and Wang and Robins (1998). The FI variance estimator for the domain mean also has a positive bias, though much smaller than that of MI. The bias in the FI variance estimator can be reduced by increasing M , but the bias of MI has little relationship to M .

All variance estimators for the variance of $\hat{\theta}_4$ are slightly negatively biased. We believe FI is slightly biased for $\hat{\theta}_4$ because, although we use the z -vector, the weights are slightly smoothed by the regression procedure. MI is known to have a small sample bias. See Kim (2002).

Table 6.2
Mean and Variance of the Point Estimators Under Setup A (5,000 Samples of Size 100)

Parameter	Imputation Scheme	Mean	Variance	Stand. Var.
Mean (θ_1)	Complete Sample	1.00	0.00570	67
	FI(3)	1.00	0.00849	100
	ABB(3)	1.00	0.00926	109
	FI(5)	1.00	0.00849	100
	ABB(5)	1.00	0.00903	106
Domain Mean (θ_2)	Complete Sample	1.14	0.02020	99
	FI(3)	1.14	0.02050	100
	ABB(3)	1.14	0.02230	109
	FI(5)	1.14	0.02040	100
	ABB(5)	1.14	0.02170	106
Pr($Y < 2$) (θ_3)	Complete Sample	0.87	0.00104	51
	FI(3)	0.87	0.00202	100
	ABB(3)	0.87	0.00228	113
	FI(5)	0.87	0.00202	100
	ABB(5)	0.87	0.00223	110
Pr($Y < 1$) (θ_4)	Complete Sample	0.50	0.00208	66
	FI(3)	0.50	0.00313	100
	ABB(3)	0.50	0.00342	109
	FI(5)	0.50	0.00313	100
	ABB(5)	0.50	0.00329	105

Table 6.3
Relative Mean, t -statistic and Relative Variance for the Variance Estimators Under Setup A
(5,000 Samples of Size 100)

Parameter	Method	Relative Mean (%)**	t -statistic*	Relative Variance (%)
Mean (θ_1)	FI(3)	100.1	0.05	5.66
	ABB(3)	99.6	-0.19	19.25
	FI(5)	100.1	0.03	5.65
	ABB(5)	98.2	-0.89	9.95
Domain Mean (θ_2)	FI(3)	115.9	7.54	13.88
	ABB(3)	127.9	12.72	28.88
	FI(5)	106.6	3.14	11.62
	ABB(5)	128.4	13.43	20.03
Pr($Y < 2$) (θ_3)	FI(3)	103.9	1.86	13.90
	ABB(3)	100.8	0.36	48.42
	FI(5)	101.7	0.82	12.07
	ABB(5)	98.5	-0.67	25.10
Pr($Y < 1$) (θ_4)	FI(3)	98.5	-0.75	4.67
	ABB(3)	96.3	-1.80	18.51
	FI(5)	97.6	-1.20	4.45
	ABB(5)	96.7	-1.65	10.17

* Statistic for hypothesis that the estimated variance is unbiased.

** Monte Carlo mean of variance estimates divided by Monte Carlo variance of estimates, in percent.

In a second set of parameters, denoted by C , the means were as follows:

Cell 1 of strata 1–25; $\mu = 0.4$

Cell 1 of strata 26–50; $\mu = 3.0$

Cell 2 of strata 1–25; $\mu = 1.6$

Cell 2 of strata 26–50; $\mu = 2.2$.

All other parameters are the same as in parameter set A. The properties of the estimators are given in Table 6.4. Both FI and MI produce unbiased estimates of the means and of the domain mean. As with parameter set A, the FI procedure is eight to twelve percent more efficient than MI for $M = 5$ and 14 to 16 percent more efficient for $M = 3$.

The assumptions required for MI variance estimation are not satisfied for parameter set C. Therefore the MI estimated

variance is seriously biased for all parameters. See Table 6.5. The bias in the MI estimated variance with $M = 5$ is about 17% for the variance of the overall mean and nearly 50% for the domain mean. The bias of the MI variance of the mean for a binomial variable is smaller than the bias for the mean of the continuous variable because the stratification effect is smaller for the binomial variable.

The properties of the estimated variances for the FI procedures are similar to those for setup A. There is a positive bias for the variance of the domain mean of about 23% for $M = 3$ and about 6% for $M = 5$.

The variance of the MI estimated variance is 2.4 to 3.5 times the variance of the FI estimated variance for $M = 5$ and 3 to 7 times for $M = 3$, demonstrating the clear superiority of the FI variance estimator for this configuration.

Table 6.4
Mean and Variance of the Point Estimators Under Setup C (5,000 Samples of Size 100)

Parameter	Imputation Scheme	Mean	Variance	Stand. Variance
Mean (θ_1)	Complete Sample	2.10	0.00500	48
	FI(3)	2.10	0.01050	100
	ABB(3)	2.10	0.01220	116
	FI(5)	2.10	0.01050	100
	ABB(5)	2.10	0.01150	110
Domain Mean (θ_2)	Complete Sample		0.02530	102
	FI(3)	2.01	0.02510	101
	ABB(3)	2.01	0.02850	115
	FI(5)	2.01	0.02480	100
	ABB(5)	2.01	0.02710	109
Pr($Y < 2$) (θ_3)	Complete Sample		0.00127	45
	FI(3)	0.45	0.00281	100
	ABB(3)	0.45	0.00322	115
	FI(5)	0.45	0.00280	100
	ABB(5)	0.45	0.00314	112
Pr($Y < 1$) (θ_4)	Complete Sample		0.00107	54
	FI(3)	0.15	0.00199	100
	ABB(3)	0.15	0.00226	114
	FI(5)	0.15	0.00199	100
	ABB(5)	0.15	0.00214	108

Table 6.5
Relative Mean, t -statistic and Relative Variance for the Variance Estimators Under Setup C (5,000 Samples of Size 100)

Parameter	Method	Relative Mean (%)	t -statistic*	Relative Variance (%)
Mean (θ_1)	FI(3)	100.9	0.41	6.42
	ABB(3)	116.7	7.31	40.14
	FI(5)	100.8	0.39	6.42
	ABB(5)	117.1	7.99	22.29
Domain Mean (θ_2)	FI(3)	122.7	10.78	16.23
	ABB(3)	144.4	19.79	46.05
	FI(5)	106.1	2.95	11.95
	ABB(5)	148.7	22.51	32.49
Pr($Y < 2$) (θ_3)	FI(3)	104.4	2.18	6.63
	ABB(3)	114.7	6.54	42.32
	FI(5)	101.8	0.89	6.42
	ABB(5)	112.1	5.74	20.67
Pr($Y < 1$) (θ_4)	FI(3)	102.3	1.13	11.08
	ABB(3)	101.3	0.58	39.14
	FI(5)	99.9	-0.04	10.05
	ABB(5)	102.2	1.04	23.60

* Statistic for hypothesis that the estimated variance is unbiased.

7. Summary

In fractional imputation, several donors are used for each missing value and each donor is given a fraction of the weight of the nonrespondent. If all donors are used, the procedure is fully efficient, under the model, for all functions of a y -vector. It is shown that the use of fractional imputation with a small number of imputations per non-respondent can give a fully efficient estimator of the mean. Estimates of other parameters, such as estimates of the cumulative distribution are nearly fully efficient.

Fractional imputation permits the construction of general purpose replicates for variance estimation. A single set of replicates can be used for variance estimation for imputed variables, variables observed on all respondents, and under model assumptions, for functions of the two types of variables. The replicates give estimates of the variances of domain means with much smaller biases than those of multiple imputation. The bias goes to zero as M increases and, in the simulation, is modest for $M = 5$. The replication variance estimator is easily implemented with replication software such as Wesvar.

Fractional imputation with a fixed number of donors per recipient is slightly more efficient for the mean than multiple imputation with the same number of donors. Fractional imputation gives variance estimates with smaller bias and much smaller variance than multiple imputation estimators with the same number of imputations.

8. Acknowledgements

This research was partially supported by a subcontract between Westat and Iowa State University under Contract No. ED-99-CO-0109 between Westat and the Department of Education and by Cooperative Agreement 13-3AEU-0-80064 between Iowa State University, the U.S. National Agricultural Statistics Service and the U.S. Bureau of the Census. We thank Jean Opsomer and Damiao Da Silva for useful comments.

References

- Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.
- Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429-440.
- Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.
- Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- Ford, B.M. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.
- Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.
- Fuller, W.A. (1996). *Introduction to Statistical Time Series*. 2nd Edition. New York: John Wiley & Sons, Inc.
- Fuller, W.A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, (Eds. R.L. Chambers and C.J. Shinner). Wiley, Chichester, England, 307-322.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics Part A – Theory and Methods*, 13, 1919-1939.
- Kim, J.K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, 89, 470-477.
- Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.
- Kim, J.K., Navarro, A. and Fuller, W.A. (2005). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, to appear.
- Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. New York: John Wiley & Sons, Inc.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-573.
- Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with applications to imputation for missing data. *Biometrika*, 82, 453-460.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- Rubin, D.B., and Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Rubin, D.B. (1987). *Multiple Imputation For Nonresponse In Surveys*. New York: John Wiley & Sons, Inc.
- Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 339-349.
- Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

- Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.
- Tollefson, M., and Fuller, W.A. (1992). Variance estimation for sampling with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.
- Wang, N., and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.
- Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of American Statistical Association*, 95, 903-915.

Variance Estimation with Hot Deck Imputation: A Simulation Study of Three Methods

J. Michael Brick, Michael E. Jones, Graham Kalton and Richard Valliant¹

Abstract

Complete data methods for estimating the variances of survey estimates are biased when some data are imputed. This paper uses simulation to compare the performance of the model-assisted, the adjusted jackknife, and the multiple imputation methods for estimating the variance of a total when missing items have been imputed using hot deck imputation. The simulation studies the properties of the variance estimates for imputed estimates of totals for the full population and for domains from a single-stage disproportionate stratified sample design when underlying assumptions, such as unbiasedness of the point estimate and item responses being randomly missing within hot deck cells, do not hold. The variance estimators for full population estimates produce confidence intervals with coverage rates near the nominal level even under modest departures from the assumptions, but this finding does not apply for the domain estimates. Coverage is most sensitive to bias in the point estimates. As the simulation demonstrates, even if an imputation method gives almost unbiased estimates for the full population, estimates for domains may be very biased.

Key Words: Adjusted jackknife; Domain estimation; Model-assisted variance estimation; Multiple imputation; Nonresponse.

1. Introduction

Imputation is frequently used in survey research to assign values for missing item responses, thereby producing complete data sets for public use or general analysis. It is well-recognized that treating imputed values as observed values results in downwardly biased variance estimates for the survey estimates. As a result, confidence intervals have lower than nominal levels. The biases in the variance estimates tend to increase with the item nonresponse rate and can be substantial when that rate is high.

Three methods of variance estimation that have been developed for use with imputed data are studied here: a model-assisted method (Särndal 1992), an adjusted jackknife method (Rao and Shao 1992), and multiple imputation (Rubin 1987). Each method has been evaluated theoretically and by simulation methods, primarily under conditions consistent with the assumptions of the methods. This paper uses simulation to compare the three methods under the same experimental conditions in which some of the assumptions required by the methods do not hold. The goal is to examine the relative performances of the methods in situations that are likely to occur in practice. Other simulation studies of variance estimation methods with imputed data have generally been more limited. Even the more extensive simulation study by Lee, Rancourt, and Särndal (2001) was based on small populations and it did not include multiple imputation.

A single-stage disproportionate stratified sample selected from a real population data set is used to evaluate these variance estimation methods in a realistic setting. The imputed values are assigned using a hot deck imputation method, one of the most popular methods of imputation in survey research. Since hot deck imputation is a form of regression imputation (Kalton and Kasprzyk 1986), restricting the simulation study to the hot deck is not a crucial feature for examining the implications for variance estimation. We study estimation for both full population and domain totals. For the domain estimates, the domain indicator is assumed to be known for all sample members.

Three different combinations of missing data mechanisms and hot deck cell formation are used in the simulations to assess the performance of the variance estimation methods under conditions that violate the assumptions of the methods to varying degrees. The three variance estimation methods we study all assume that data are randomly missing in each hot deck cell and the model-assisted (MA) and multiple imputation (MI) methods also assume that a simple model with common mean and variance holds in each cell. Studying the robustness of the variance estimation methods is an important feature of the simulation because in practice the assumptions underlying the methods will almost never be fully satisfied.

The next section briefly describes three variance estimation methods with hot deck imputed data. The third section outlines the study population, the sample design used in the simulations, and the methods used to generate the missing

1. J. Michael Brick, Michael E. Jones and Graham Kalton, Westat, 1650 Research Boulevard, Rockville, MD 20850; Richard Valliant, University of Michigan, 1218 Lefrak Hall, College Park, MD 20742.

data and implement the hot deck imputations. The fourth section gives the results of the simulations. The last section gives some conclusions about the methods and their applicability.

2. Description of the Variance Estimation Methods

We denote the full sample by A , the subset that responds to an item by A_R , and the subset that does not respond by A_M . For the imputations the units are divided into hot deck cells indexed by $g=1, \dots, G$, where the subset of n_{Rg} respondents in cell g is A_{Rg} , and the subset of non-respondents is A_{Mg} . For each unit with a missing value, the hot deck method consists of randomly selecting a respondent from within the same hot deck cell to be the donor of the imputed value.

With hot deck imputation, donors are often selected within a cell by simple random sampling with replacement (srswr), by simple random sampling without replacement, or by sampling with probabilities proportional to the survey weights with replacement (ppswr). Since the simulation results obtained using the srswr and the ppswr methods are very similar, only the results for the ppswr method—termed the weighted hot deck—are presented here. The imputed estimator of a population total is $\hat{\theta}_I = \sum_{i \in A_R} w_i y_i + \sum_{i \in A_M} w_i y_i^*$, where w_i is the survey weight, y_i is the reported value and y_i^* is the imputed value for unit i in the nonrespondent set.

2.1 Model-Assisted Variance Estimation

The model-assisted (MA) approach with hot deck imputation assumes that data are randomly missing within the hot deck cells and that a model for the generation of the y_i 's holds. A natural model for use with hot deck imputation is that the y_i 's are independently and identically generated within the hot deck cells, i.e., $y_{gi} \stackrel{\text{iid}}{\sim} (\mu_g, \sigma_g^2)$ for cell g . Inferences from the model-assisted approach depend on the validity of the model assumptions.

Särndal (1992) decomposed the total variance of the imputed estimator into three components denoted by V_{SAM} , V_{IMP} , and V_{MIX} . The estimators used for these components in the simulations are those given in Brick, Kalton, and Kim (2004). The MA variance estimator is the sum of the component estimates: $\hat{V}_{\text{MA}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}}$. The \hat{V}_{IMP} and \hat{V}_{MIX} estimators require an estimator of the element variance in each hot deck cell. Since the simulations showed little difference between weighted and unweighted estimators only the weighted estimator of σ_g^2 is discussed, that is $\hat{\sigma}_g^2 = n_{Rg} (n_{Rg} - 1)^{-1} \sum_{i \in A_{Rg}} w_i (y_i - \bar{y}_{Rg})^2 \times (\sum_{i \in A_{Rg}} w_i)^{-1}$, with $\bar{y}_{Rg} = \sum_{i \in A_{Rg}} w_i y_i / (\sum_{i \in A_{Rg}} w_i)$.

2.2 Adjusted Jackknife Variance Estimation

The Rao and Shao (1992) adjusted jackknife (AJ) variance estimator for a stratified sample with imputations and ignorable finite population correction factors (fpc 's) is

$$\hat{V}(\hat{\theta}_I) = \sum_{h=1}^L \sum_{k=1}^{n_h} \frac{n_h - 1}{n_h} (\hat{\theta}_{Ih}^{(k)} - \hat{\theta}_I)^2,$$

where n_h is the number sampled in stratum h ,

$$\hat{\theta}_{Ih}^{(k)} = \sum_{g=1}^G \left\{ \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)} y_{hi} + \sum_{(hj) \in A_{Mg}} w_{hj}^{(k)} (y_{hj}^* + \hat{y}_{Rg}^{(k)} - \bar{y}_{Rg}) \right\}$$

is the adjusted estimator when unit k is omitted,

$$\hat{y}_{Rg}^{(k)} = \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)} y_{hi} / \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)},$$

$$\bar{y}_{Rg} = \sum_{(hi) \in A_{Rg}} w_{hi} y_{hi} / \sum_{(hi) \in A_{Rg}} w_{hi},$$

$w_{hi}^{(k)}$ is the weight for unit hi adjusted to account for the omission of unit k . The notation $(hi) \in B$ denotes unit i in stratum h is part of set B . This procedure requires the computation of $\sum n_h$ replicate estimates, $\hat{\theta}_{Ih}^{(k)}$. A commonly used strategy to reduce the computations is to combine units into variance strata (e.g., see Rust and Rao 1996). Let h^* denote a combined variance stratum and k a group of sample units within the combined stratum. All sampled units are assigned to one of the groups. Then, the grouped adjusted jackknife variance estimator is

$$\hat{V}_{AJ} = \sum_{h^*} \sum_{k=1}^{n_{h^*}} \frac{n_{h^*} - 1}{n_{h^*}} (\hat{\theta}_{Ih^*}^{(k)} - \hat{\theta}_I)^2,$$

where n_{h^*} is the number of sample units in combined variance stratum h^* , $n_{h^*(k)}$ is the number of units retained in stratum h^* when units in group k are deleted and, corresponding to $\hat{\theta}_{Ih^*}^{(k)}$, $\hat{\theta}_{Ih^*}^{(k)}$ is the adjusted imputed estimate for the full population when units in group k in stratum h^* are deleted. The retained units from design stratum h that are in combined variance stratum h^* are assigned replicate weights of $w_{hi}^{(k)} = n_{h^*} (n_{h^*(k)})^{-1} w_{hi}$.

The AJ method assumes a uniform response probability model within each hot deck cell but, unlike the MA method, it does not require distributional assumptions. Under the uniform response probability model without distributional assumptions, a weighted hot deck is needed to produce unbiased imputed estimates.

In developing the theory for the AJ method, Rao and Shao (1992) assume that fpc 's are ignorable. However, the fpc 's are not negligible in some strata in the simulations, ranging from about 0.05 to 0.24. Shao and Steel (1999) and Lee, Rancourt, and Särndal (1995) provide methods for accounting for nonnegligible fpc 's. The Lee, Rancourt, and Särndal (1995) fpc adjustment was applied in the simulations because of its ease of implementation. Without the

fpc adjustment, the AJ variance estimator substantially overestimated the variances in the simulations.

2.3 Multiple Imputation

Multiple imputation (MI) is described in detail in Rubin (1987) and Little and Rubin (2002). The summary here relates to its application with hot deck imputation. As with the model-assisted approach, within the hot deck cells responses are assumed to be missing randomly and the *y*'s are assumed to be independent random variables with a common mean and variance. For each unit that has a missing value, *M* values are imputed, creating *M* completed data sets.

To avoid underestimation of variances with the MI method, the hot deck method needs to be modified. Rubin and Schenker (1986) proposed the approximate Bayesian bootstrap (ABB) for simple random sampling with hot deck imputation for use with the MI method. The ABB was modified for the simulations to accommodate sampling donors by ppswr. In the simulations a donor pool for the ABB was created in each cell by selecting respondents with replacement with probabilities proportional to *w_i*. (There is no literature that discusses the application of ABB methods with unequal weights. In hindsight, an unweighted ABB might have been preferable. The use of an unweighted ABB with a ppswr hot deck yields unbiased point estimates of population totals under the response probability model).

3. Design of the Simulation Study

3.1 Description of the Study Population and Sample Design

The sampling frame for the simulations is a subset of the file of public school districts extracted from the 1999–2000 Common Core of Data (CCD) compiled by the U.S. National Center for Education Statistics. The final frame consists of 11,941 districts.

The sample design used in the simulations is a stratified simple random sample of 1,020 school districts. Twelve strata were created by cross-classifying four categories of number of students (district size) by three categories of the percentage of students at or below the poverty level (poverty status). The strata and number of districts in the frame are given in Table 1. The table also gives the stratum sample sizes and sampling rates used in the simulations.

The table also contains the stratum means and standard deviations for the two study variables, the number of students in the district and the number of districts that include pre-kindergarten as the lowest grade. These study variables were chosen because they are typical of many estimates computed from this type of design.

In addition to the full population estimates we computed the two study estimates for two domains, defined as districts located in the Northeast region and those in nonmetropolitan areas. The means for these domains are substantially different from the full population means for both study variables.

3.2 Missing Data Mechanisms and Imputation Methods

By construction, information on the two study variables is available for all districts in the sampling frame. To create missing values, response indicators were assigned to sampled units within “response cells”. In some cases the response cells are the sampling strata, termed STR cells, whereas in other cases they are what are termed HD cells. The HD cells were defined by the cross-classification of four geographic regions and a fourfold categorization of the number of full time equivalent teachers in the district. The HD cells are somewhat correlated with the sampling strata, but each cell contains units from more than one stratum.

Table 1
Stratum Definitions, Population Counts, Sample Sizes, Sampling Rates, Means and Standard Deviations of Number of Students and Proportions of Districts with Pre-Kindergarten

Stratum	District size	Poverty status	<i>N_h</i>	<i>n_h</i>	Sampling rate	Number of students		Proportion with pre-kindergarten
						Mean	Std. dev.	
1	1	1	615	32	0.0520	270.0	155.0	0.44
2	1	2	1,147	59	0.0514	263.3	175.0	0.49
3	1	3	1,292	66	0.0511	243.5	142.5	0.49
4	2	1	1,720	111	0.0645	1,607.2	837.0	0.44
5	2	2	2,305	149	0.0646	1,429.7	784.1	0.52
6	2	3	1,893	122	0.0644	1,427.8	788.8	0.63
7	3	1	692	75	0.1084	4,695.3	1,360.6	0.35
8	3	2	579	63	0.1088	4,728.5	1,365.0	0.51
9	3	3	527	57	0.1082	4,591.8	1,380.3	0.63
10	4	1	342	83	0.2427	16,003.4	12,670.2	0.51
11	4	2	449	110	0.2450	17,577.3	14,246.7	0.58
12	4	3	380	93	0.2447	19,331.8	16,142.7	0.68
Total			11,941	1,020		3,237.9	6,770.5	0.52

Within a given response cell, sampled units were assigned at random to be missing or nonmissing at a specified rate. For each type of response cell, three schemes for assigning rates of missingness were chosen. In two of the schemes, the rates of missingness varied across the response cells, whereas in the other scheme the rate was constant across the cells.

The simulations were conducted by first drawing a stratified simple random sample using the stratum sample sizes in Table 1. Once the sample was selected, response status (respondent/nonrespondent) was randomly assigned to each sampled unit according to the given response scheme. For the MA and AJ methods, the weighted hot deck imputation procedures described earlier were used to impute for missing values. For the MI method, a donor pool was first created using the weighted ABB, and weighted hot decks were then used to impute for each of the $M=5$ imputed data sets. The estimated total numbers of students and districts with pre-kindergarten were computed for the simulated sample with imputed values, and variance estimates were computed for these estimates using the three variance estimation methods. (If the estimated variance could not be computed in a particular simulation run or the sample size in a cell was less than 2, then that sample was deleted. The maximum number of deleted samples across all the simulations of 10,000 runs each was 2 for the MA method and 28 for the AJ (only one run had 28 AJ samples deleted; the next largest number was 3). The AJ method was based on three combined variance strata and 40 groups of units per stratum for a total of 120 replicates. The three combined strata, formed from strata having about the same *fpc*, consisted of strata 1–6, 7–9, and 10–12. As a check of the grouping, we verified that the grouped jackknife variance procedure gave essentially the same average variance estimates and confidence interval coverage rates as the ungrouped jackknife in the case of complete response. The entire process was repeated 10,000 times for each response scheme.

A feature of the design of the simulation is that the means for the two domains considered often differ substantially from the full population means by strata and HD cells. A key point for the domain estimates is that imputations were made by selecting donors from all the respondents in a hot deck cell, without specifically recognizing the domain as might be done in practice for some domains. After imputations were made for the full sample, the estimated total for a domain was estimated by $\hat{\theta}_I = \sum_{i \in A_R} \delta_i w_i y_i + \sum_{j \in A_M} \delta_j w_j y_j^*$ where $\delta_i = 1$ if unit i is in the domain and 0 if not.

Three of the four possible combinations of response mechanism (STR or HD cells) and hot deck cell formation (STR or HD cells) were studied in the simulations. We refer to these combinations as STR/STR, HD/HD, and STR/HD,

where the first set of letters identifies the response mechanism and the second set identifies the type of hot deck cell. The three sets of response rates were 0.2 to 0.6 spaced evenly across the response cells, a constant 0.7 in all cells, and 0.6 to 0.9 spread evenly across the cells. The three combinations of response/hot deck cells with the three sets of response rates generated nine separate simulation schemes for each estimate.

3.3 Assumptions for Models of Response and Population Structure

There are two models involved in the simulations. The population model assumes that the y values within each hot deck cell are independent and have the same expected value. The response model assumes that there is a uniform response probability within each hot deck cell. If both models hold, then the use of either an unweighted or a weighted hot deck will lead to an unbiased estimate of the overall population total. However, if only the response model is assumed, then the use of a weighted hot deck is needed to produce an unbiased estimate of the overall population total. Since the weighted hot deck is used in the simulations, only the response probability model needs to be satisfied for unbiased point estimation of the overall population total. The response probability model holds for all the STR/STR and HD/HD combinations and for the STR/HD combination with a constant response rate; however, it does not hold for the other two STR/HD combinations. The AJ theory for variance estimation of population totals was developed assuming only the response probability model. The MA and MI theories assume that both models hold.

Reliance on only the response probability model and the weighted hot deck to produce unbiased estimates of population totals does not in general extend to estimates of domain totals. When domains cut across hot deck cells, it is necessary to invoke a population model that assumes that the expected value of the domain values is the same as that of the nondomain values in each hot deck cell. However, if the hot deck cells are defined such that each domain comprises the full population in a subset of the hot deck cells, then the situation for point and variance estimation is the same as stated above for overall population totals.

The simulation schemes were generally constructed so that the hot deck cells do not incorporate the domains in order to reflect the practical consideration that it is essentially impossible to incorporate all domains in an imputation scheme. Specifically, in the simulations the districts in the Northeast (NE) region and districts in nonmetropolitan statistical areas (NMSA) are unrelated to the stratum definitions in Table 1 (which are used as hot deck cells in some cases). Also, districts in the NMSA domain can be found in all HD cells. However, the NE

domain is a subset of four of the HD cells. Thus, the definition of the HD cells is more consistent with estimating NE domain totals than NMSA domain totals.

3.4 Summary Statistics

The relative bias of a point estimate is estimated by $relbias(\hat{\theta}_I) = bias(\hat{\theta}_I) / \theta_N$, where $bias(\hat{\theta}_I) = \sum_s (\hat{\theta}_{Is} - \theta_N) / 10,000$, $\hat{\theta}_{Is}$ is the estimate from sample s , and θ_N is the finite population parameter. The empirical variance of $\hat{\theta}_I$ is $Var(\hat{\theta}_I) = \sum_s (\hat{\theta}_{Is} - \bar{\theta}_I)^2 / 10,000$, where $\bar{\theta}_I = \sum_s \hat{\theta}_{Is} / 10,000$. The average variance estimate for a particular method is $v = \sum_s v_s / 10,000$, where v_s is the estimated variance for simulation run s .

The percentages of intervals that include θ_N are based on the nominal 95 percent confidence intervals ($\hat{\theta}_I \pm t\hat{V}^{1/2}$) computed for each of the 10,000 simulations for each simulation scheme. An issue to consider here is the precision of the variance estimates from a disproportionate stratified sample design and its impact on whether normal approximation or t intervals should be used to calculate confidence intervals. We found that the use of the t -distribution did not have a substantial effect for most cases with the MA and AJ methods, and we have therefore used a multiplier of 1.96 for confidence intervals based on these methods. Rubin and Schenker (1986) suggest using a t -distribution with λ degrees of freedom for confidence intervals with the MI method, where

$$\lambda = (M - 1) \left(1 + \frac{M}{M + 1} \frac{U}{B} \right)^2.$$

Since using 1.96 with the MI method yielded intervals that had severe undercoverage, the t -distribution with λ degrees of freedom is used for the MI confidence intervals.

4. Simulation Results

This section presents the main results from the simulations, beginning with the performance of the three methods of variance estimation for estimates from the full population, followed by the results for the domain estimates. Key outcomes are summarized here graphically, but tables with full details are available in Brick, Jones, Kalton, and Valliant (2004).

4.1 Full Population Estimates

Figure 1 shows the results of the simulations for estimating the total number of students and the number of districts offering pre-kindergarten from the 10,000 samples for each of the nine simulation schemes. The figure gives the relative bias of the imputed estimator, the average variance estimate as a percentage of the empirical variance, and the confidence interval coverage rate.

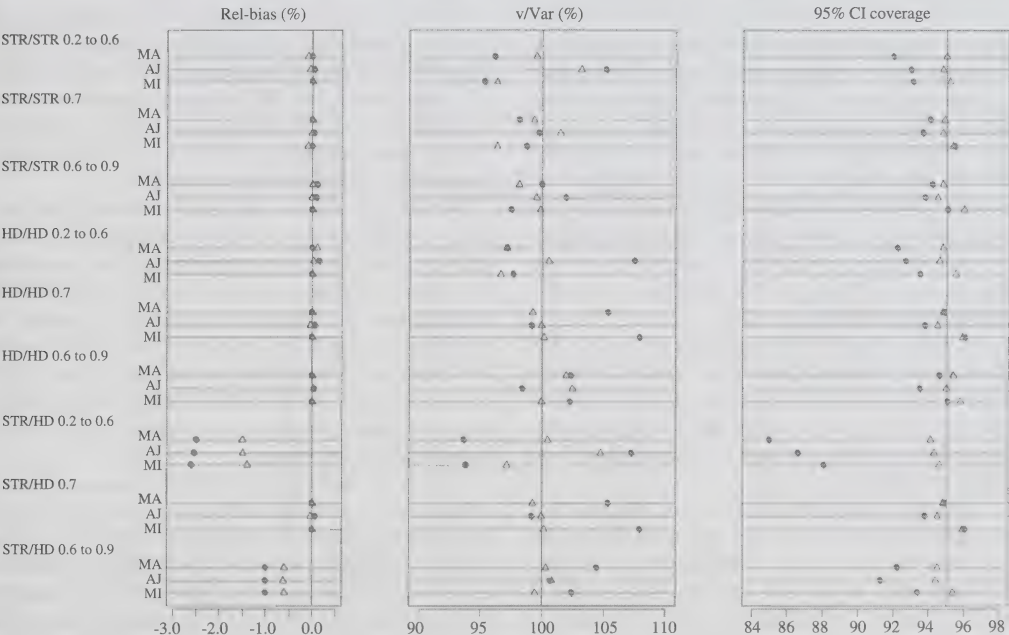


Figure 1. Relative biases, variance ratios, and 95% confidence interval coverage for number of students (•) and number of districts with pre-kindergarten (Δ).

The point estimates are theoretically unbiased with weighted hot deck imputation if all units in a hot deck cell have the same response probability. As noted earlier, this condition holds for the STR/STR and HD/HD combinations and also for the STR/HD combination with a uniform overall response probability. The graph of relative biases in Figure 1 is consistent with this theoretical result within the bounds of simulation error. While the relative biases of the point estimates in the other two STR/HD schemes are small (always less than 3%), they still may be important if the standard errors of the estimates are also small. Cochran (1977, page 12) shows that when the ratio of the bias to the standard error is relatively large, then the coverage rate can be much lower than the nominal level. For the full population estimates with this sample size the ratios never exceed 0.4, but much larger ratios occur for domain estimates, as discussed later.

The graph of the ratios of the average variance estimates to the empirical variances (v/Var in the figures) for the three methods shows that these estimates have relatively small biases in most cases, within a range of plus or minus 8 percent around the simulated true variance. While the ratios for all the methods vary across the nine schemes, the MI ratios are slightly more variable than the other two.

A primary reason for computing variances is to produce confidence intervals. The right-hand panel in Figure 1 shows that the coverage rates for the confidence intervals for the estimates are generally close to the nominal 95 percent level, especially for the pre-kindergarten statistic. The coverage rates for both statistics and all the methods and schemes are between 91% and 96%, with the exception of the number of students for the STR/HD 0.2 to 0.6 scheme. The coverage rates of 88% or less for all three methods in this case, with its extremely high rate of nonresponse, are due to the relatively large bias in the point estimate. Overall, all three variance estimation methods produce confidence intervals with coverages that are vast improvements over those for intervals based on naïve variance estimates (Brick *et al.* 2004).

The confidence interval coverage rates for the MA and AJ methods are essentially equivalent. The MI coverage rates are generally slightly greater than those for the MA and AJ methods. The MI coverage rates are slightly closer to the nominal level for the number of students. Most of the differences are small.

For all three variance estimation methods, the upper and lower confidence interval coverage rates were similar. For the number of students, which is a highly skewed variable, the coverage rates in the two tails are unequal due to correlation between the estimated total and the standard error estimates. The asymmetric tail coverages are also associated with lower overall coverage rates.

The MA and AJ methods yield confidence intervals that have nearly the same average length across the schemes and variables. Because the MI method uses t -distribution values, its intervals range from 10 to 20 percent longer than the MA and AJ intervals when the response rates are low. With the higher response rates, the MI intervals range from about the same to 5 percent longer than the intervals from the two other methods. The MI confidence intervals could, of course, be shortened by increasing M (Rubin 1987, Chapter 4), even though $M = 5$ is typical for applications.

4.2 Domain Estimates

Estimating characteristics for domains that are not explicitly incorporated in the imputation scheme can be problematic when the missing data rate is not trivial. Kalton and Kasprzyk (1986) and Rubin (1996) along with many others have discussed this point and urged the inclusion of as many variables as possible in the imputation process. However, given the many preplanned and ad hoc domain analyses that are carried out with survey data, it is unrealistic to assume that all domains can be accounted for in an imputation scheme. For this reason, the design of the simulations intentionally did not include the domains explicitly in the definition of the hot deck cells. In the case of multiple imputation, issues of variance estimation for domain estimates have received much attention (*e.g.*, Fay 1992; Meng 1994; Rubin 1996).

In the simulations we estimate the totals for two domains: school districts in the NE and those in NMSA. Figures 2 and 3 present the results of the simulations for the NE domain and for the NMSA domain, respectively, in the same format as used before. Note that the scales for Figures 2 and 3 differ from each other and are very different from those used for the full population estimates.

For the NE domain, the point estimates have large positive biases for the STR/STR combinations. Hot deck cells based on STR are not related to region, and, as a result, NE districts with missing data have donors from other regions, which have different characteristics. In contrast, the inclusion of region in the construction of the HD imputation cells removes the bias of the point estimates in the HD/HD combinations and the STR/HD combination with uniform overall response probability, and reduces the bias in the other STR/HD combinations.

All three methods of variance estimation require unbiased point estimates and theory for the methods does not provide guidance on how the methods will perform under the conditions we study. The variance estimates are approximately unbiased for all three variance estimation methods when the domain point estimates are unbiased or have only small biases. However, Figure 2 shows that for the STR/STR combination, where the point estimates are

seriously biased, the variance estimates usually overestimate the empirical variances.

Figure 2 shows that the coverage rates for the HD/HD and STR/HD schemes—for which the point estimates have no or small relative biases—are between 92 percent and 96 percent for all but one of these schemes and variance estimation methods. The exception is the STR/HD combination with response rates between 0.2 and 0.6, which has coverage rates as low as 86 percent for the number of students.

For the STR/STR schemes, Figure 2 shows that all the methods tend to cover at greater than the nominal level for the number of students and less than the nominal level for the number of districts with pre-kindergarten. The difference in the coverage rates for the two variables is due to the sizes of the relative bias of the point estimates and of the variance estimates.

Turning to the NMSA domain estimates in Figure 3, note that metropolitan status is not explicitly included in the

definitions of either STR or HD, although it is clearly correlated with size and, thus, with STR. The point estimates for the number of students in the NMSA domain for all the schemes have substantial positive biases. The MA confidence intervals consistently cover at the nominal level or higher, primarily due to the extreme positive biases of the variance estimates. The AJ intervals cover at close to the nominal level for the HD/HD and STR/HD schemes, but undercover in the three STR/STR schemes. The patterns for the MI coverages are similar to those of the AJ, except that the MI intervals appreciably undercover in the HD/HD scheme with 0.2 to 0.6 response rates.

The point estimates of the number of districts with pre-kindergarten in the NMSA domain have moderate negative relative biases for all nine schemes. The confidence intervals for all three methods of variance estimation are close to the nominal level, without the overcoverage found in the NE domain estimates.

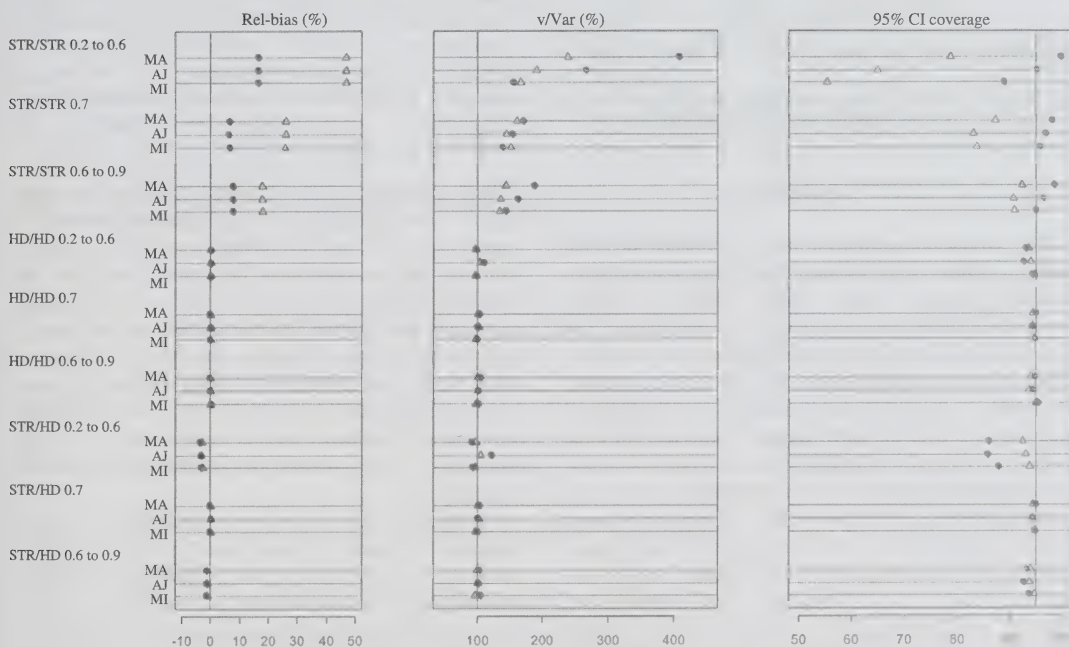


Figure 2. Relative biases, variance ratios, and 95% confidence interval coverage for number of students (•) and number of districts with pre-kindergarten (Δ) in the Northeast.

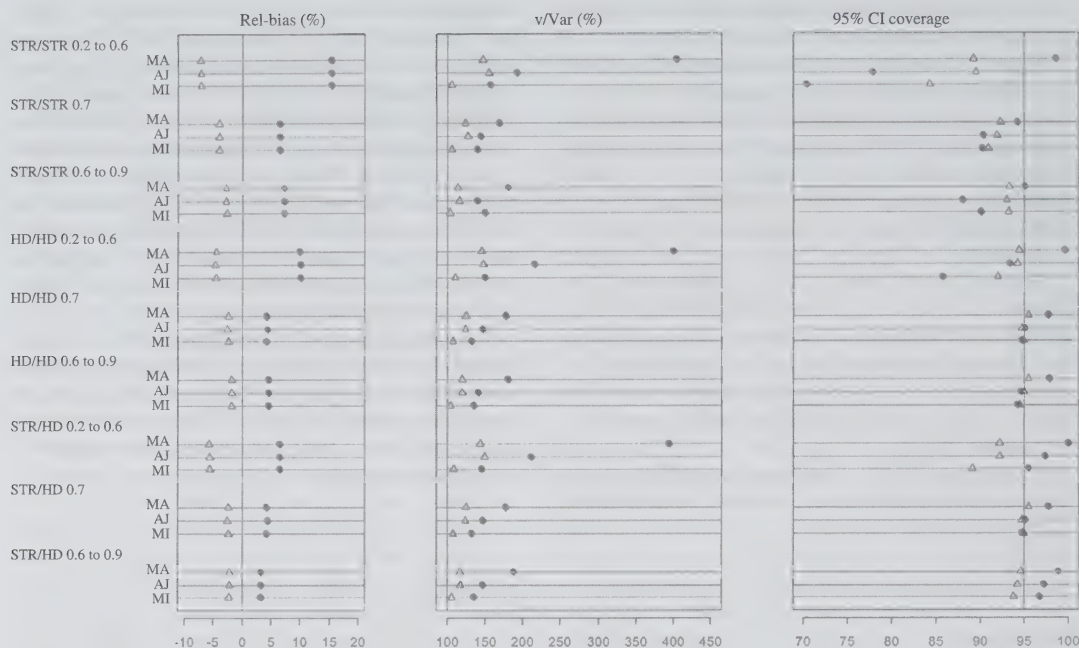


Figure 3. Relative biases, variance ratios, and 95% confidence interval coverage for number of students (●) and number of districts with pre-kindergarten (Δ) in nonmetropolitan areas.

5. Conclusions

The simulations examined the performance of three variance estimators for imputed totals from a single-stage stratified sample design under different response mechanisms with weighted hot deck imputation. The circumstances reflected what can be expected in practice in the sense that the assumptions of the methods were violated in different ways. All three methods were substantial improvements over the naïve variance estimator. All three methods performed very well with unbiased point estimates. When the point estimates had large biases, none of the methods produced confidence intervals with the nominal coverage levels. Poor coverage rates for biased point estimates are not unexpected since the same result holds with no missing data. When the point estimates had relatively small biases, the actual coverage rates for the three variance estimation methods sometimes exceeded and sometimes fell short of the nominal levels. In this case the tendency of all three methods to overestimate the variance often resulted in coverage rates close to the nominal level. Low response rates were associated with undercoverage, largely due to the greater biases in the point estimates.

The differences in the coverage rates of the three methods were generally too small and inconsistent to support claims that any one method is superior in general. With very low response rates, the average lengths of the confidence intervals for the MI method were appreciably longer than those for the MA and AJ methods, but using a larger number of sets of imputations with the MI method would rectify that problem. It should, however, be noted that these simulations only address single stage sampling. Differences in confidence interval lengths between methods may exist in cluster samples. This possibility awaits further investigation.

The results of this study give practitioners of hot deck imputation empirical evidence that all of the variance estimation methods perform well in single stage samples provided that the point estimate is unbiased, even when other assumptions are violated. Estimates for domains that are not taken into account in the imputation scheme are susceptible to large biases. When the point estimates are seriously biased, the methods may produce confidence intervals that cover at far less than the nominal rate. Analysts of imputed data sets should examine whether the imputation method that has been used is likely to give approximately unbiased estimates, especially for domain

estimates. If not, they may need to re-impute the missing items to give less biased point estimates. Advice to imputers to take advantage of as many explanatory variables as feasible in the imputation process is not new, but the evidence from the simulations demonstrates its importance.

Acknowledgements

The authors would like to thank the National Center for Education Statistics, Institute for Education Sciences for supporting this research, and in particular Marilyn Seastrom. We also would like to thank the referees for their constructive comments.

References

- Brick, J.M., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57-66.
- Brick, J.M., Jones, M., Kalton, G. and Valliant, R. (2004). A simulation study of three methods of variance estimation with hot deck imputation for stratified samples. Prepared under contract No. RN95127001 to the National Center for Education Statistics. Rockville, MD: Westat, Inc.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons Inc.
- Fay, R.E. (1992). When are imputations from multiple imputation valid. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Lee, H., Rancourt, E. and Särndal, C.-E. (1995). Jackknife variance estimation for data with imputed values. *Proceedings of the Statistical Society of Canada Survey Methods Section*, 111-115.
- Lee, H., Rancourt, E. and Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A Little), Chapter 21, New York: John Wiley & Sons Inc.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons Inc.
- Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input. (With discussion). *Statistical Science*, 9, 538-573.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons Inc.
- Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.
- Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with nonignorable nonresponse. *Journal of the American Statistical Association*, 81, 361-374.
- Rust, K., and Rao, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medicine*, 5, 381-397.
- Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite estimation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Does Weighting for Nonresponse Increase the Variance of Survey Means?

Roderick J. Little and Sonya Vartivarian¹

Abstract

Nonresponse weighting is a common method for handling unit nonresponse in surveys. The method is aimed at reducing nonresponse bias, and it is often accompanied by an increase in variance. Hence, the efficacy of weighting adjustments is often seen as a bias-variance trade-off. This view is an oversimplification – nonresponse weighting can in fact lead to a reduction in variance as well as bias. A covariate for a weighting adjustment must have two characteristics to reduce nonresponse bias – it needs to be related to the probability of response, and it needs to be related to the survey outcome. If the latter is true, then weighting can reduce, not increase, sampling variance. A detailed analysis of bias and variance is provided in the setting of weighting for an estimate of a survey mean based on adjustment cells. The analysis suggests that the most important feature of variables for inclusion in weighting adjustments is that they are predictive of survey outcomes; prediction of the propensity to respond is a secondary, though useful, goal. Empirical estimates of root mean squared error for assessing when weighting is effective are proposed and evaluated in a simulation study. A simple composite estimator based on the empirical root mean squared error yields some gains over the weighted estimator in the simulations.

Key Words: Missing data; Nonresponse adjustment; Sampling weights; Survey nonresponse.

1. Introduction

In most surveys, some individuals provide no information because of noncontact or refusal to respond (*unit nonresponse*). The most common method of adjustment for unit nonresponse is weighting, where respondents and nonrespondents are classified into adjustment cells based on covariate information known for all units in the sample, and a nonresponse weight is computed for cases in a cell proportional to the inverse of the response rate in the cell. These weights often multiply the sample weight, and the overall weight is normalized to sum to the number of respondents in the sample. A good overview of nonresponse weighting is Oh and Scheuren (1983). A related approach to nonresponse weighting is post-stratification (Holt and Smith 1979), which applies when the distribution of the population over adjustment cells is available from external sources, such as a Census. The weight is then proportional to the ratio of the population count in a cell to the number of respondents in that cell.

Nonresponse weighting is primarily viewed as a device for reducing bias from unit nonresponse. This role of weighting is analogous to the role of sampling weights, and is related to the design unbiasedness property of the Horvitz-Thompson estimator of the total (Horvitz and Thompson 1952), which weights units by the inverse of their selection probabilities. Nonresponse weighting can be viewed as a natural extension of this idea, where included units are weighted by the inverse of their inclusion

probabilities, estimated as the product of the probability of selection and the probability of response given selection; the inverse of the latter probability is the nonresponse weight. Modelers have argued that weighting for bias adjustment is not necessary for models where the weights are not associated with the survey outcomes, but in practice few are willing to make such a strong assumption.

Sampling weights reduce bias at the expense of increased variance, if the outcome has a constant variance. Given the analogy of nonresponse weights with sampling weights, it seems plausible that nonresponse weighting also reduces bias at the expense of an increase in the variance of survey estimates. The idea of a bias-variance trade-off arises in discussions of nonresponse weighting adjustments (Kalton and Kasprzyk 1986, Kish 1992, Little, Lewitzky, Heeringa, Lepkowski and Kessler 1997). Kish (1992) presents a simple formula for the proportional increase in variance from weighting, say L , under the assumption that the variance of the observations is approximately constant:

$$L = cv^2, \quad (1)$$

where cv is the coefficient of variation of the respondent weights.

Equation (1) is a good approximation when the adjustment cell variable is weakly associated with the survey outcome. However, since it approximates variance rather than mean squared error, it does not measure the potential nonresponse bias reduction that is the main objective of weighting, and it does not apply to outcomes

1. Roderick J. Little, University of Michigan, U.S.A. E-mail: rlittle@umich.edu; Sonya Vartivarian, Mathematica Policy Research, Inc. 600 Maryland Ave SW, Suite 550, Washington, D.C. 20024-2512. E-mail: SVartivarian@Mathematica-MPR.com.

that are associated with the adjustment cell variable, where nonresponse weighting can in fact reduce the variance. The fact that nonresponse weighting can reduce variance is implicit in the formulae in Oh and Scheuren (1983), and is noted in Little (1986) when adjustment cells are created using predictive mean stratification. It is also seen in the related method of post-stratification for nonresponse adjustment (Holt and Smith 1979).

Variability of the weights per se does not necessarily translate into estimates with high variance: an estimate with a high value of L can have a smaller variance than an estimate with a small value of L , as is shown in the simulations in section 3. Also, the situations where nonresponse weighting is most effective in reducing bias are precisely the situations where the weighting tends to reduce, not increase, variance, and Equation (1) does not apply. This differs from the case of sampling weights, and is related to "super-efficiency" that can result when weights are estimated from the sample rather than fixed constants; see, for example, Robins, Rotnitzky and Zhao (1994).

We propose a simple refinement of Equation (1), namely Equation (14) below, that captures both bias and variance components whether or not the adjustment cell variable is associated with the outcome, and hence is a more accurate gauge of the value of weighting the estimates, and of alternative adjustment cell variables. In multipurpose surveys with many outcomes, the standard approach is to apply the same nonresponse weighting adjustment to all the variables, with the implicit assumption that the value of nonresponse bias reduction for some variables outweighs the potential variance increase for others. Our empirical estimate of mean squared error allows a simple refinement of this strategy, namely to restrict nonresponse weighting to the subset of variables for which nonresponse weighting reduces the estimated mean squared error. This composite strategy is assessed in the simulation study in section 3, and shows some gains over weighting all the outcomes. As noted in section 4, there are alternative approaches that have even better statistical properties, but these lead to different weights for each variable and hence are more cumbersome to implement and explain to survey users.

2. Nonresponse Weighting Adjustments for a Mean

Suppose a sample of n units is selected. We consider inference for the population mean of a survey variable Y subject to nonresponse. To keep things simple and focused on the nonresponse adjustment question, we assume that units are selected by simple random sampling. The points made here about nonresponse adjustments also apply in

general to complex designs, although the technical details become more complicated.

We assume that respondents and nonrespondents can be classified into C adjustment cells based on a covariate X . Let M be a missing-data indicator taking the value 0 for respondents and 1 for nonrespondents. Let n_{mc} be the number of sampled individuals with $M = m$, $X = c$, $m = 0, 1$; $c = 1, \dots, C$, $n_{+c} = n_{0c} + n_{1c}$ denote the number of sampled individuals in cell c , $n_0 = \sum_{c=1}^C n_{0c}$ and $n_1 = \sum_{c=1}^C n_{1c}$ the total number of respondents and nonrespondents, and $p_c = n_{+c}/n$, $p_{0c} = n_{0c}/n_0$ the proportions of sampled and responding cases in cell c . We compare two estimates of the population mean μ of Y , the unweighted mean

$$\bar{y}_0 = \sum_{c=1}^C p_{0c} \bar{y}_{0c}, \quad (2)$$

where \bar{y}_{0c} is the respondent mean in cell c , and the weighted mean

$$\bar{y}_w = \sum_{c=1}^C p_c \bar{y}_{0c} = \sum_{c=1}^C w_c p_{0c} \bar{y}_{0c}, \quad (3)$$

which weights respondents in cell c by the inverse of the response rate $w_c = p_c / p_{0c}$. The estimator (3) can be viewed as a special case of a regression estimator, where missing values are imputed by the regression of Y on indicators for the adjustment cells. We compare the bias and mean squared error of (2) and (3) under the following model, which captures the important features of the problem. We suppose that conditional on the sample size n , the sampled cases have a multinomial distribution over the $(C \times 2)$ contingency table based on the classification of M and X , with cell probabilities

$$\Pr(M = 0, X = c) = \phi \pi_{0c}; \Pr(M = 1, X = c) = (1 - \phi) \pi_{1c},$$

where $\phi = \Pr(M = 0)$ is the marginal probability of response. The conditional distribution of X given $M = 0$ and n_0 is multinomial with cell probabilities $\Pr(X = c | M = 0) = \pi_{0c}$, and the marginal distribution of X given n is multinomial with index n and cell probabilities

$$\Pr(X = c) = \phi \pi_{0c} + (1 - \phi) \pi_{1c} = \pi_c,$$

say. We assume that the conditional distribution of Y given $M = m$, $X = c$ has mean μ_{mc} and constant variance σ^2 . The mean of Y for respondents and nonrespondents are

$$\mu_0 = \sum_{c=1}^C \pi_{0c} \mu_{0c}, \mu_1 = \sum_{c=1}^C \pi_{1c} \mu_{1c},$$

respectively, and the overall mean of Y is $\mu = \phi \mu_0 + (1 - \phi) \mu_1$.

Under this model, the conditional mean and variance of \bar{y}_w given $\{p_c\}$ are respectively $\sum_{c=1}^C p_c \mu_{0c}$ and $\sigma^2 \sum_{c=1}^C p_c^2 / n_{0c}$. Hence the bias of \bar{y}_w is

$$b(\bar{y}_w) = \sum_{c=1}^C \pi_c (\mu_{0c} - \mu_c),$$

where π_c and μ_c are the population proportion and mean of Y in cell c . This can be written as

$$b(\bar{y}_w) = \tilde{\mu}_0 - \mu, \quad (4)$$

where $\tilde{\mu}_0 = \sum_{c=1}^C \pi_c \mu_{0c}$ is the respondent mean “adjusted” for the covariates, and $\mu = \sum_{c=1}^C \pi_c \mu_c$ is the true population mean of Y . The variance of \bar{y}_w is the sum of the expected value of the conditional variance and the variance of its conditional expectation, and is approximately

$$V(\bar{y}_w) = (1 + \lambda) \sigma^2 / n_0 + \sum_{c=1}^C \pi_c (\mu_{0c} - \tilde{\mu}_0)^2 / n, \quad (5)$$

where $\lambda = \sum_{c=1}^C \pi_{0c} (\pi_c / \pi_{0c} - 1)^2$ is the population analog of the variance of the nonresponse weights $\{w_c\}$, which is the same as L in Equation (1) since the weights are scaled to average to one. The formula for the variance of the weighted mean in Oh and Scheuren (1983), derived under the quasi-randomization perspective, reduces to (5) when the within-cell variance is assumed constant, and finite population corrections and terms of order $1/n^2$ are ignored. The mean squared error of \bar{y}_w is thus

$$\text{mse}(\bar{y}_w) = b^2(\bar{y}_w) + V(\bar{y}_w). \quad (6)$$

The mean squared error of the unweighted mean (2) is

$$\text{mse}(\bar{y}_0) = b^2(\bar{y}_0) + V(\bar{y}_0), \quad (7)$$

where:

$$b(\bar{y}_0) = b(\bar{y}_w) + \mu_0 - \tilde{\mu}_0, \quad (8)$$

is the bias and

$$V(\bar{y}_0) = \sigma^2 / n_0 + \sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2 / n_0, \quad (9)$$

is the variance. Hence the difference (say Δ) in mean squared errors is

$$\Delta = \text{mse}(\bar{y}_0) - \text{mse}(\bar{y}_w) = B + V_1 - V_2, \text{ where}$$

$$B = (\mu_0 - \tilde{\mu}_0)^2 + 2(\mu_0 - \tilde{\mu}_0)(\tilde{\mu}_0 - \mu),$$

$$V_1 = \sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2 / n_0 - \sum_{c=1}^C \pi_c (\mu_{0c} - \tilde{\mu}_0)^2 / n,$$

$$V_2 = \lambda \sigma^2 / n_0 \quad (10)$$

Equation (10) and its detailed interpretation provide the main results of the paper; note that positive terms in (10) favor the weighted estimator \bar{y}_w .

- (a) The first term B represents the impact on MSE of bias reduction from adjustment on the covariates. It is order one and increasingly dominates the MSE as the sample size increases. If $\mu \leq \tilde{\mu}_0 < \mu_0$ or $\mu_0 < \tilde{\mu}_0 \leq \mu$, then weighting has reduced the bias of the respondent

mean, and both of the components of B are positive. In particular, if the missing data are missing at random (Rubin 1976, Little and Rubin 2002), in the sense that respondents are a random sample of the sampled cases in each cell c , then $\tilde{\mu}_0 = \mu$ and weighting eliminates the bias of the unweighted mean. The bias adjustment is

$$\mu_0 - \tilde{\mu}_0 \approx \sum_{c=1}^C \pi_{0c} (1 - w_c) (\mu_{0c} - \mu_0),$$

ignoring differences between the weights and their expectations. This is zero to $O(1)$ if either non-response is unrelated to the adjustment cells (in which case $w_c \approx 1$ for all c , or the outcome is unrelated to the adjustment cells (in which case $\mu_{0c} \approx \mu_0$ for all c). Thus a substantial bias reduction requires adjustment cell variables that are related both to nonresponse and to the outcome of interest, a fact that has been noted by several authors. It is often believed that conditioning on observed characteristics of nonrespondents will reduce bias, but note that this is not guaranteed; it is possible for the adjusted mean to be further on average from the true mean than the unadjusted mean, in which case weighting makes the bias worse.

- (b) The effect of weighting on the variance is represented by $V_1 - V_2$.
- (c) For outcomes Y that are unrelated to the adjustment cells, $\mu_{0c} = \mu_0$ for all c , $V_1 = 0$, and weighting increases the variance, since V_2 is positive. The variance part of equation (10) then reduces to the population version of Kish's formula (1). Adjustment cell variables that are good predictors of nonresponse hurt rather than help in this situation, since they increase the variance of the weights without any reduction in bias; but there is no bias-variance trade-off for these outcomes, since there is no bias reduction.
- (d) If the adjustment cell variable X is unrelated to non-response, then λ is $O(1/n)$ and hence V_2 has a lower order of variability than V_1 . The term V_1 tends to be positive, since $\sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2 \approx \sum_{c=1}^C \pi_{0c} (\mu_{0c} - \tilde{\mu}_0)^2$, and the divisor n in the second term is larger than the divisor n_0 in the first term. Thus weighting in this case tends to have no impact on the bias, but reduces variance to the extent that X is a good predictor of the outcome. This contradicts the notion that weighting increases variance. The above-mentioned “super-efficiency” that results from estimating non-response weights from the sample is seen by the fact that if the data are missing completely at random, then the “true” nonresponse weight is a constant for all responding units. Hence weighting by “true” weights

leads to (2), which is less efficient than weighting by the “estimated” weights, which leads to (3).

- (e) If the adjustment cell variable is a good predictor of the outcome and also predictive of nonresponse, then V_2 is again small because of the reduced residual variance σ^2 , and V_1 is generally positive by a similar argument to (d). The term $\sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2$ may deviate more from $\sum_{c=1}^C \pi_c (\mu_{0c} - \mu_0)^2$ because the weights are less alike, but this difference could be positive or negative, and the different divisors seem more likely to determine the sign and size of V_1 . Thus, weighting tends to reduce both bias and variance in this case.
- (f) Equation (9) can be applied to the case of post-stratification on population counts, by letting n represent the population size rather than the sample size. Assuming a large population, the second term in V_1 essentially vanishes, increasing the potential for variance reduction when the variables forming the post-strata are predictive of the outcome. This finding replicates previous results on post-stratification (Holt and Smith 1979; Little 1993).

A simple qualitative summary of the results (a) – (f) of section 2 is shown in Table 1, which indicates the direction of bias and variance when the associations between the adjustment cells and the outcome and missing indicator are high or low. Clearly, weighting is only effective for outcomes that are associated with the adjustment cell variable, since otherwise it increases the variance with no compensating reduction in bias. For outcomes that are associated with the adjustment cell variable, weighting increases precision, and also reduces bias if the adjustment cell variable is related to nonresponse.

Table 1

Effect of Weighting Adjustments on Bias and Variance of a Mean, by Strength of Association of the Adjustment Cell Variables with Nonresponse and Outcome

Association with nonresponse	Association with outcome	
	Low	High
Low	Cell 1	Cell 3
	Bias: ---	Bias: ---
	Var: ---	Var: ↓
High	Cell 2	Cell 4
	Bias: ---	Bias: ↓
	Var: ↑	Var: ↓

It is useful to have estimates of the MSE of \bar{y}_0 and \bar{y}_w that can be computed from the observed data. Let $s_{0c}^2 = \sum_{i \in c} (y_i - \bar{y}_{0c})^2 / (n_{0c} - 1)$ denote the sample variance of respondents in cell c , $s^2 = \sum_{c=1}^C (n_{0c} - 1) s_{0c}^2 / (n_{0c} - C)$ the pooled within-cell variance, and $s_0^2 = \sum_{i=1}^{n_0} (y_i - \bar{y}_0)^2 / (n_0 - 1)$, the total sample variance of the respondent

values. We use the following approximately unbiased expressions, under the assumption that the data are MAR:

$$\text{mse}(\bar{y}_0) = \hat{B}^2(\bar{y}_0) + \hat{V}(\bar{y}_0), \quad (11)$$

where $\hat{V}(\bar{y}_0) = s_0^2 / n_0$ and

$$\hat{B}^2(\bar{y}_0) = \max\{0, (\bar{y}_w - \bar{y}_0)^2 - V_d\}$$

$$V_d = (n_1 / n)^2 \left[\sum_{c=1}^C p_{1c} (\bar{y}_{0c} - \bar{y}_0^{(1)})^2 / n_1 + \sum_{c=1}^C p_{0c} (\bar{y}_{0c} - \bar{y}_0)^2 / n_0 + s^2 \sum_{c=1}^C (p_{1c} - p_{0c})^2 / n_{0c} \right], \quad (12)$$

where $\bar{y}_0^{(1)} = \sum_{c=1}^C p_{1c} \bar{y}_{0c}$, and V_d estimates the variance of $(\bar{y}_w - \bar{y}_0)$ and is included in (12) as a bias adjustment for $(\bar{y}_w - \bar{y}_0)^2$ as an estimate of $B^2(\bar{y}_0)$, similar to that in Little *et al.* (1997). Also

$$\text{mse}(\bar{y}_w) = \hat{V}(\bar{y}_w) = (1 + L) s^2 / n_0 + \sum_{c=1}^C p_c (\bar{y}_{0c} - \bar{y}_w)^2 / n. \quad (13)$$

Subtracting (11) from (13), the difference in MSE's of \bar{y}_w and \bar{y}_0 is then estimated by

$$D = L s^2 / n_0 - (s_0^2 - s^2) / n_0 + \sum_{c=1}^C p_c (\bar{y}_{0c} - \bar{y}_w)^2 / n - \hat{B}^2(\bar{y}_0). \quad (14)$$

This is our proposed refinement of (1), which is represented by the leading term on the right side of (14).

3. Simulation Study

We include simulations to illustrate the bias and variance of the weighted and unweighted mean for sets of parameters representing each cell in Table 1. We also compare the analytic MSE approximations in Equations (6) and (7) and their sample-based estimates (11) and (13) with the empirical MSE over repeated samples.

3.1 Superpopulation Parameters

The simulation set-up for the joint distribution of X and M is described in Table 2. The sample is approximately uniformly distributed across the adjustment cell variable X , which has $C = 10$ cells. Two marginal response rates are chosen, 70%, corresponding to a typical survey value, and 52%, a more extreme value to accentuate differences in methods. Three distributions of M given X are simulated to model high, medium and low association.

The simulated distributions of the outcome Y given $M = m, X = c$ are shown in Table 3. These all have the form

$$[Y | M = m, X = c] \sim N(\beta_0 + \beta_1 X, \sigma^2).$$

Table 2
Percent of Sample Cases in Adjustment Cell *X* and Missingness Cell *M*

a. Overall Response Rate = 52%

	Association Between <i>M</i> and <i>X</i>	<i>X</i>	1	2	3	4	5	6	7	8	9	10
1.	High	<i>M</i> = 0	0.55	1.00	4.01	4.52	5.04	5.55	6.06	6.58	9.14	9.96
		<i>M</i> = 1	8.69	9.00	6.01	5.53	5.04	4.54	4.04	3.54	1.02	0.20
2.	Medium	<i>M</i> = 0	2.77	3.50	4.01	4.52	5.04	5.55	6.06	6.58	7.11	7.62
		<i>M</i> = 1	6.47	6.50	6.01	5.53	5.04	4.54	4.04	3.54	3.05	2.54
3.	Low	<i>M</i> = 0	4.62	5.15	5.21	5.28	5.34	5.40	5.45	5.52	5.58	5.64
		<i>M</i> = 1	4.62	4.85	4.81	4.77	4.73	4.69	4.65	4.60	4.57	4.52

b. Overall Response Rate = 70%

	Association Between <i>M</i> and <i>X</i>	<i>X</i>	1	2	3	4	5	6	7	8	9	10
1.	High	<i>M</i> = 0	0.55	3.00	6.51	7.04	7.55	8.07	8.59	9.11	9.64	9.96
		<i>M</i> = 1	8.69	7.00	3.51	3.02	2.52	2.02	1.52	1.01	0.51	0.20
2.	Medium	<i>M</i> = 0	4.44	5.30	5.81	6.33	6.85	7.37	7.88	8.40	8.93	9.45
		<i>M</i> = 1	4.80	4.70	4.21	3.72	3.22	2.72	2.22	1.72	1.22	0.71
3.	Low	<i>M</i> = 0	6.19	6.85	6.91	6.98	7.05	7.11	7.17	7.24	7.31	7.37
		<i>M</i> = 1	3.05	3.15	3.11	3.07	3.02	2.98	2.93	2.88	2.84	2.79

Table 3
Parameters for $[Y | M = m, X = c] \sim N(\beta_0 + \beta_1 c, \sigma^2)$

Association Between <i>Y</i> and <i>X</i>	β_1	σ^2	ρ^2
1. High	4.75	46	≈ 0.80
2. Medium	3.70	122	≈ 0.48
3. Low	0.00	234	0.00

Three sets of values of (β_1, σ^2) are simulated to model high, medium and low associations between *Y* and *X*. The intercept β_0 is chosen so that the overall mean of *Y* is $\mu = 26.3625$ for each scenario.

A thousand replicate samples of size $n = 400$ and $n = 2,000$ were simulated for each combination of parameters in Tables 2 and 3. Samples where $n_{0c} = 0$ for any *c* were excluded, since the weighted estimate cannot be computed; in practice some cells would probably be pooled in such cases. The numbers of excluded simulations are shown in Table 4.

Table 4
Numbers of Replicates Excluded Because of Cell with no Respondents

Association of <i>M</i> and <i>X</i>	Association of <i>Y</i> and <i>X</i>	Response Rate	
		52%	70%
High	High	134	113
	Medium	120	117
	Low	131	104
Medium	Low	1	0

3.2 Comparisons of Bias, Variance and Root Mean Squared Error, and their Estimates

Summaries of empirical bias and root MSE's (RMSE's) are reported in Table 5. The empirical RMSE's of the weighted mean can be compared with the following estimates, which are displayed in Table 5, averaged over the 1,000 replicates: The estimated RMSE based on Kish's rule of thumb Equation (1), namely:

$$mse_{Kish}(\bar{y}_w) = (1 + L)s_y^2 / n_0,$$
$$\text{where } s_y^2 = \sum_{i=1}^{n_0} (y_i - \bar{y}_0)^2 / (n_0 - 1); \tag{15}$$

The analytical RMSE from Equations (6) and (7); and the estimated RMSE from Equations (11) and (13).

Following the suggestion of Oh and Scheuren (1983), we include in the last two columns of Table 5 the average empirical bias and RMSE of a composite mean that chooses between \bar{y}_w and \bar{y}_0 , picking the estimate with a lower sample-based estimate of the MSE. The empirical bias relative to the population parameter is reported for all estimators. We also include the bias and RMSE of the mean before deletion of cases due to nonresponse.

Table 5a shows results for simulations with a response rate of 52%. Rows are labeled according to the four cells in Table 1, with medium and high associations combined. For each row, the lower of the RMSE's for the unweighted and weighted respondent means is bolded, indicating superiority for the corresponding method.

The first four rows of Table 5a correspond to cell 4 in Table 1, with medium/high association between *Y* and *X* and

medium/high association between M and X . In these cases \bar{y}_w has much lower RMSE than \bar{y}_0 , reflecting substantial bias of \bar{y}_0 that is removed by the weighting.

The next two rows of Table 5a corresponding to cell 3 of Table 1, with medium/high association between Y and X and low association between M and X . In these cases \bar{y}_0 is no longer seriously biased, but \bar{y}_w has improved precision, particularly when the association of Y and X is high. These are cases where the variance is reduced, not increased, by weighting. The analytic estimates of RMSE and sample-based estimates are close to the empirical RMSE estimates, while Kish's rule of thumb overestimates the RMSE, as predicted by the theory in section 2.

The next two rows of Table 5a correspond to cell 2 of Table 1, where the association between Y and X is low and the association between M and X is medium or high. In these cases, \bar{y}_w has higher MSE than \bar{y}_0 . These cases illustrate situations where the weighting increases variance, with no compensating reduction in bias. The last row corresponds to cell 1 of Table 1, with low associations between M and X and between Y and X . The unweighted mean has lower RMSE in these cases, but the increase in RMSE from weighting is negligible. For the last three rows of Table 5a, RMSE's from Kish's rule of thumb are similar

to those from the analytical formula in section 2 and empirical estimates based on this formulae, and all these estimates are close to the empirical RMSE.

The last two columns of Table 5a show empirical bias and RMSE of the composite method that chooses \bar{y}_w or \bar{y}_0 based on the estimated RMSE. For the simulations in the first 6 rows, the composite estimator is the same as \bar{y}_w , and hence detects and removes the bias of the unweighted mean. For simulations in cell 1 (the last row) the composite estimator performs like \bar{y}_w or \bar{y}_0 , as expected since \bar{y}_w and \bar{y}_0 perform similarly in this case. For simulations in cell 2 that are not favorable to weighting, the composite estimator has lower RMSE than \bar{y}_w , but considerably higher than that of \bar{y}_0 , suggesting that for the conditions of this simulation the empirical MSE affords limited ability to pick the better estimator in individual samples.

Nevertheless, the composite estimator is the best overall estimator of the three considered in this simulation.

Table 5b shows results for the 70% response rate. The pattern of results is very similar to that of Table 5a. As expected, differences between the methods are smaller, although they remain substantial in many rows of the table.

Table 5a

Summaries of Estimators Based on 1,000 Replicate Samples for $C = 10$ Adjustment Cells, Restricted to Sample Replicates with $n_{0c} > 0$ for all c . Response Rate of 52%. Values are Multiplied by 1,000

Association with Adjustment Cells Based on X				Unweighted Mean				Weighted Mean				Before Deletion Mean		Composite Mean		
Cell	(M, X)	(Y, X)	n	emp. bias	emp. rmse	analytical rmse ¹	est. rmse ²	emp. bias	emp. rmse	Kish rmse ³	analytical rmse ⁴	est. rmse ⁵	emp. bias	emp. rmse	emp. bias	emp. rmse
4	High	High	400	6,955	7,024	7,055	6,974	0	1,057	1,410	956	988	-38	795	0	1,057
			2,000	7,008	7,020	7,006	7,015	-2	424	608	427	434	12	342	-2	424
4	High	Medium	400	5,376	5,471	5,536	5,404	-33	1,264	1,510	1,216	1,297	-21	776	-33	1,264
			2,000	5,424	5,441	5,466	5,466	-41	561	650	545	559	-30	338	-41	561
4	Medium	High	400	3,664	3,794	3,809	3,754	-4	816	1,071	835	842	6	741	-4	816
			2,000	3,703	3,731	3,700	3,712	7	369	473	373	374	4	337	7	369
4	Medium	Medium	400	2,838	3,006	3,042	2,991	-18	938	1,095	954	970	-9	747	-18	938
			2,000	2,864	2,900	2,898	2,893	-2	426	483	426	428	6	335	-2	426
3	Low	High	400	476	1,148	1,113	1,178	40	823	1,050	823	828	30	764	40	823
			2,000	376	587	614	595	-11	361	465	368	368	-3	333	-11	361
3	Low	Medium	400	350	1,106	1,095	1,134	13	927	1,063	925	939	-16	762	13	927
			2,000	287	565	563	559	-20	429	470	413	414	-22	353	-20	429
2	High	Low(0)	400	56	1,070	1,056	1,275	96	1,658	1,613	1,518	1,631	28	793	83	1,410
			2,000	-11	464	473	567	-26	698	698	679	699	-19	337	-25	620
2	Medium	Low(0)	400	9	1,042	1,053	1,077	-27	1,122	1,112	1,097	1,125	21	772	-12	1,074
			2,000	-4	474	471	480	-11	491	491	491	493	11	340	-9	481
1	Low	Low(0)	400	-30	1,038	1,050	1,055	-30	1,053	1,064	1,050	1,076	-30	752	-30	1,040
			2,000	-2	472	469	469	-1	474	470	469	471	-8	343	-1	472

¹ Computed using Equation (7)

² Computed using Equation (11)

³ Computed using Equation (15)

⁴ Computed using Equation (6)

⁵ Computed using Equation (13)

Table 5b
Summaries of Estimators based on 1,000 Replicate Samples for C = 10 Adjustment Cells, Restricted to Sample Replicates with $n_{0c} > 0$ for all c. Response Rate of 70%. Values are Multiplied by 1,000

Association with Adjustment				Unweighted				Weighted				Before Deletion		Composite		
Cells based on X				Mean				Mean				Mean		Mean		
Cell	(M, X)	(Y, X)	n	emp. bias	emp. rmse	analytical rmse ⁶	est. rmse ⁷	emp. bias	emp. rmse	Kish rmse ⁸	analytical rmse ⁹	est. rmse ¹⁰	emp. bias	emp. rmse	emp. bias	emp. rmse
4	High	High	400	4,692	4,810	4,893	4,860	-133	1,129	1,192	889	894	-129	998	-133	1,129
			2,000	4,827	4,841	4,839	4,854	-20	400	529	398	405	-5	334	-20	400
4	High	Medium	400	3,581	3,716	3,855	3,733	-133	1,266	1,250	1,075	1,097	-128	917	-127	1,284
			2,000	3,763	3,784	3,778	3,777	-9	501	554	481	490	11	343	-9	501
4	Medium	High	400	2,666	2,812	2,878	2,837	-58	803	910	794	796	-49	772	-58	803
			2,000	2,732	2,760	2,767	2,761	-6	353	406	355	355	-9	333	-6	353
4	Medium	Medium	400	2,104	2,282	2,315	2,291	-28	833	924	854	861	-43	751	-28	833
			2,000	2,146	2,180	2,170	2,165	13	370	411	382	382	10	334	13	370
3	Low	High	400	217	906	954	980	-81	797	911	790	793	-77	771	-81	797
			2,000	312	513	506	502	2	365	405	353	353	4	349	2	365
3	Low	Medium	400	251	922	942	960	15	804	916	845	852	26	727	15	804
			2,000	224	454	472	471	-14	370	408	378	379	-15	327	-14	370
2	High	Low(0)	400	0	952	915	1,131	35	1,445	1,349	1,298	1,358	1	807	26	1,292
			2,000	-11	416	409	485	-41	608	598	580	599	-4	347	-31	535
2	Medium	Low(0)	400	22	911	910	920	24	942	936	930	946	2	757	21	925
			2,000	23	418	407	411	20	425	416	416	417	15	344	19	420
1	Low	Low(0)	400	1	914	914	912	2	917	916	914	926	-5	751	1	914
			2,000	4	402	408	408	4	403	409	408	410	6	331	4	402

⁶ Computed using Equation (7)
⁷ Computed using Equation (11)
⁸ Computed using Equation (15)
⁹ Computed using Equation (6)
¹⁰ Computed using Equation (13)

4. Discussion

The results in sections 2 and 3 have important implications for the use of weighting as an adjustment tool for unit nonresponse. Surveys often have many outcome variables, and the same weights are usually applied to all these outcomes. The analysis of section 2 and simulations in section 3 suggests that improved results might be obtained by estimating the MSE of the weighted and unweighted mean and confining weighting to cases where this relationship is substantial. A more sophisticated approach is to apply random-effects models to shrink the weights, with more shrinkage for outcomes that are not strongly related to the covariates (e.g., Elliott and Little 2000). A flexible alternative to this approach is imputation based on prediction models, since these models allow for interval-scaled as well as categorical predictors, and allow interactions to be dropped to incorporate more main effects. Multiple imputation (Rubin 1987) can be used to propagate uncertainty.

When there is substantial covariate information, one attractive approach to generalizing weighting class adjustments is to create a propensity score for each respondent based on a logistic regression of the nonresponse indicator on the covariates, and then create adjustment cells based on this score. Propensity score methods were originally

developed in the context of matching cases and controls in observational studies (Rosenbaum and Rubin 1983), but are now quite commonly applied in the setting of unit nonresponse (Little 1986; Czajka, Hirabayashi, Little and Rubin 1987; Ezzati and Khare 1992). The analysis here suggests that for this approach to be productive, the propensity score has to be predictive of the outcomes. Vartivarian and Little (2002) consider adjustment cells based on joint classification by the response propensity and summary predictors of the outcomes, to exploit residual associations between the covariates and the outcome after adjusting for the propensity score. The requirement that adjustment cell variables predict the outcomes lends support to this approach.

The analysis presented here might be extended in a number of ways. Second order terms in the variance are ignored here, which if included would penalize weighting adjustments based on a large number of small adjustment cells. Finite population corrections could be included, although it seems unlikely that they would affect the main conclusions. It would be of interest to see to what extent the results can be generalized to complex sample designs involving clustering and stratification. Also, careful analysis of the bias and variance implications of nonresponse weighting on statistics other than means, such as subclass means or regression coefficients, would be worthwhile. We

expect it to be important that adjustment cell variables predict the outcome in many of these analyses too, but other points of interest may emerge.

Acknowledgements

This research is supported by grant SES-0106914 from the National Science Foundation. We thank an associate editor and three referees for useful comments on earlier drafts.

References

- Czajka, J.L., Hirabayashi, S.M., Little, R.J.A. and Rubin, D.B. (1987). Evaluation of a new procedure for estimating income aggregates from advance data. In *Statistics of Income and Related Administrative Record Research: 1986-1987*, U.S. Department of the Treasury, 109-136.
- Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.
- Ezzati, T., and Khare, M. (1992). Nonresponse adjustments in a National Health Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 339-344.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47, 663-685.
- Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical, A*, 142, 33-46.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Kish, L. (1992). Weighting for unequal P_i . *Journal of Official Statistics*, 8, 183-200.
- Little, R.J.A. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.
- Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- Little, R.J.A., Lewitzky, S., Heeringa, S., Lepkowski, J. and Kessler, R.C. (1997). An assessment of weighting methodology for the national comorbidity study. *American Journal of Epidemiology*, 146, 439-449.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley & Sons, Inc.
- Oh, H.L., and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*, 2, Theory and Bibliographies, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), Academic Press, New York, 143-184.
- Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.
- Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.
- Vartivarian, S., and Little, R.J.A. (2002). On the formation of weighting adjustment cells for unit nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Variance-Covariance Functions for Domain Means of Ordinal Survey Items

Alistair James O'Malley and Alan Mark Zaslavsky¹

Abstract

Estimates of a sampling variance-covariance matrix are required in many statistical analyses, particularly for multilevel analysis. In univariate problems, functions relating the variance to the mean have been used to obtain variance estimates, pooling information across units or variables. We present variance and correlation functions for multivariate means of ordinal survey items, both for complete data and for data with structured non-response. Methods are also developed for assessing model fit, and for computing composite estimators that combine direct and model-based predictions. Survey data from the Consumer Assessments of Health Plans Study (CAHPS®) illustrate the application of the methodology.

Key Words: Variance function; Correlation function; Hierarchical model; Ordinal response; Nonresponse; Skip pattern.

1. Introduction

Survey data are often used to obtain measures for comparisons across estimation domains. In our motivating example, surveys are conducted to elicit reports on experiences with health plans (entities administering health care) from enrolled members; similarly a survey might assess schools by administering tests to a sample of students.

An essential part of the analysis of survey data is the calculation of sampling variances, or the sampling-covariance matrix of a multivariate estimator. The standard survey sampling approach is to compute variances directly for each estimator in each domain. Direct variance estimates may be unstable when the number of respondents to an item is small because the sample size for a domain is small, because the item is applicable to only a fraction of respondents (such as users of specialized equipment in health surveys), or because we are interested in means for a small subgroup (such as those with chronic illnesses).

By modeling variance estimates as functions of the unit (domain) means, we can pool information across units to obtain more stable estimates. Although modeling may introduce bias, for small units this is offset by the reduction in sampling variation. One may also consider generalizing variance estimates across items in addition to or instead of domains. This will be appropriate when there are groups of items for which the same mean-variance relationship is likely to hold. However, when there are many more domains than items, the greatest potential gain is from generalizing across domains rather than across items.

A *Generalized Variance Function* (GVF) is a mathematical model describing the relationship between the

variance or relative variance of a survey estimator and its expectation. When multiple estimates are produced from the same sample, Wolter (1985, chapter 5) proposes the model

$$V/M^2 = \theta_0 + \theta_1/M,$$

where M and V denote the expected value and variance of the estimator respectively. Such a form might be suitable for variables such as income or wealth for which a nearly constant coefficient of variation might be plausible because the mean and standard deviation are proportional to the length of the reference period. Modeling the coefficient of variation is thus most suited to situations where the variables are similar in content but have different scales with unrestricted ranges (e.g., income collected monthly and yearly). In our problem the items are ordinal and so a model of the coefficient of variation is not a natural choice. Other proposed GVs also have simple forms (Woodruff 1992; Otto and Bell 1995).

If a suitable GVF can be found, it can simplify calculations and make variance estimates more stable. Furthermore, summarizing sampling variance estimates in the form of a function also facilitates presentation of large volumes of statistics (Wolter 1985, pages 201-202). Finally, modeling variances as functions of means facilitates iterative re-estimation of sampling variances in hierarchical modeling. In practice the decision to use variance functions in a hierarchical modeling context depends on the goodness of the fit of the GVF; only with a sufficiently good fit is use of the GVF worthwhile.

Past work on GVFs is relatively sparse. Wolter (1985, chapter 5) gave an overview but provided only a few references, as did Valliant, Dorfman and Royall (2000,

1. Alistair James O'Malley and Alan Mark Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899, U.S.A. E-mail: omalley@hcp.med.harvard.edu and zaslavsk@hcp.med.harvard.edu.

pages 344–348). Valliant (1992a, 1992b) used GVF's to smooth time-dependent indices in time series analysis. Woodruff (1992) used GVF's for variance estimation of employment change in the Current Employment Survey, and Wolter (1985, pages 208–217) illustrates the use of GVF's on data from the Current Population Survey. GVF's are also used in the National Health Interview Survey (Valliant *et al.* 2000, page 344).

Huff, Eltinge, and Gershunskaya (2002) and Cho, Eltinge, Gershunskaya and Huff (2002) considered GVF's for the United States Current Employment Survey and Consumer Expenditure Survey. Eltinge (2002) uses GVF's to estimate a full sampling covariance matrix when samples are too small to produce stable estimates for all areas, estimating the components of the mean squared error (MSE) of the GVF model. Otto and Bell (1995) fit GVF's to median income, per capita income, and age-group poverty rates in the Current Population Survey, assuming an autoregressive dependence between rates over time and a Wishart distribution for the sampling covariance matrices.

Our research extends previous research on GVF's in four directions. First, we use the GVF to generalize across domains rather than items. Thus, we do not assume that different items have the same GVF, although it might be reasonable to fit models of the same form for items with similar response categories. Second, we develop GVF's for the full covariance matrix, which must be estimated for joint inference on multiple outcomes. Thirdly, we focus on the relationship between means and variances of items with the ordinal response formats often used in survey questionnaires, rather than on homoscedastic continuous responses. Finally, we explicitly allow for patterns of nonresponse due to structured skip patterns. While structured item non-response can be ignored (except for its effect on sample size) in univariate estimation, it must be considered explicitly to model bivariate relationships because it affects the sampling covariance of item means. Furthermore, because the number of responses varies across items, we cannot model the sampling covariances using a Wishart distribution, which has only a single parameter for sample size.

We first describe direct estimation of variances and covariances, including the case when data are missing due to skip patterns. In section 3 we introduce models for generalized variance and covariance functions (GVCFs) and lay out our strategies for model fitting and evaluation and for combining direct estimates and model predictions. In section 4, we apply our methods to a major health care survey. In section 5, we conclude by describing applications and extensions of our methods.

2. Direct Estimates of Sampling Variances of Domain Means

We index observations by domain h , items (indices i and j), and respondents (indices k and l); $y_{h,ik}$ and $r_{h,ik}$ denote the outcome and response indicator of subject k in domain h on item i . We suppress the index for item when referring to all items for a respondent or domain, and have no need for the subscript for respondent when discussing the means, variances, and correlations of items.

Direct estimation of the sampling covariance matrix of domain means (henceforth, "variance estimation") begins by expressing the means as functions of totals of the outcomes and response indicators. We replace $y_{h,ik}$ with 0 for missing observations so that totals are defined in the presence of skip patterns. Following the notation of Särndal, Swenson and Wretman (1992, pages 24–28; 36–42), let U_h and S_h describe the population and sample respectively for the h^{th} domain, $Y_{h,i} = \sum_{U_h} y_{h,ik}$, $R_{h,i} = \sum_{U_h} r_{h,ik}$, $\hat{Y}_{h,i} = \sum_{S_h} \tilde{y}_{h,ik}$, and $\hat{R}_{h,i} = \sum_{S_h} \tilde{r}_{h,ik}$, where $\tilde{y}_{h,ik} = y_{h,ik} / \pi_{h,k}$, $\tilde{r}_{h,ik} = r_{h,ik} / \pi_{h,k}$, and $\pi_{h,k} = \text{pr}(k \in S_h)$.

The vector of mean outcomes for the population of elements within domain h is

$$M_h = f(Y_h, R_h) = \left(\frac{Y_{h,1}}{R_{h,1}}, \dots, \frac{Y_{h,I}}{R_{h,I}} \right),$$

where $Y_h = (Y_{h,1}, \dots, Y_{h,I})$ and $R_h = (R_{h,1}, \dots, R_{h,I})$. An estimator is

$$f(\hat{Y}_h, \hat{R}_h) = \left(\frac{\hat{Y}_{h,1}}{\hat{R}_{h,1}}, \dots, \frac{\hat{Y}_{h,I}}{\hat{R}_{h,I}} \right).$$

A first order Taylor series expansion of $f(\hat{Y}_h, \hat{R}_h)$ about $f(Y_h, R_h)$ produces the approximation

$$\text{var}(f(\hat{Y}_h, \hat{R}_h)) \approx V_h = f'(Y_h, R_h) \text{var}(\hat{Y}_h, \hat{R}_h) f'(Y_h, R_h)^T,$$

where $f'(Y_h, R_h)$ is the Jacobian of $f(Y_h, R_h)$. Often it is computationally easier to first calculate $u_{h,k} = f'(Y_h, R_h)z_{h,k}$, where $z_{h,k} = (y_{h,k}, r_{h,k})$, and then evaluate the variance as

$$\begin{aligned} V_h &= \text{var} \left(\sum_{S_h} \tilde{u}_{h,k} \right) \\ &= \text{var} \left(\sum_{U_h} \tilde{u}_{h,k} I_{h,k} \right) \\ &= \sum_{k,l \in U_h} \Delta_{h,kl} \tilde{u}_{h,k} \tilde{u}_{h,l}^T, \end{aligned}$$

where $I_{h,k} = 1$ if $k \in S_h$ (indicating that the k^{th} member of domain h is sampled) and 0 otherwise, $\Delta_{h,kl} = \pi_{h,kl} - \pi_{h,k} \pi_{h,l}$, and $\pi_{h,kl} = \text{pr}(k, l \in S_h)$. An estimator for V_h is

$$\hat{V}_h = \sum_{k,l \in S_h} \tilde{\Delta}_{h,kl} \tilde{u}_{h,k} \tilde{u}_{h,l}^T, \quad (1)$$

where $\tilde{\Delta}_{h,kl} = \Delta_{h,kl} / \pi_{h,kl}$.

To describe evaluation of \hat{V}_h one need only consider one diagonal (*i.e.*, variance) element and one off-diagonal (*i.e.*, covariance) element. The sub-matrix of the Jacobian formed by the i^{th} and j^{th} items is given by

$$f'(Y_h, R_h) = \begin{pmatrix} \frac{1}{R_{h,i}} & 0 & -\frac{Y_{h,i}}{R_{h,i}^2} & 0 \\ 0 & \frac{1}{R_{h,j}} & 0 & -\frac{Y_{h,j}}{R_{h,j}^2} \end{pmatrix}.$$

For, $z_{h,k} = (y_{h,ik}, y_{h,jk}, r_{h,ik}, r_{h,jk})$, it follows that

$$u_{h,k} = f'(Y_h, R_h) z_{h,k} = \begin{pmatrix} \frac{1}{R_{h,i}} (y_{h,ik} - M_{h,i} r_{h,ik}) \\ \frac{1}{R_{h,j}} (y_{h,jk} - M_{h,j} r_{h,jk}) \end{pmatrix},$$

where $M_{h,i} = Y_{h,i} / R_{h,i}$ is the mean outcome of the i^{th} item in domain h . Hence,

$$\hat{V}_{h,ii} = \frac{1}{R_{h,i}^2} \sum_{k,l \in S_h} \tilde{\Delta}_{h,kl} (\tilde{y}_{h,ik} - M_{h,i} \tilde{r}_{h,ik})(\tilde{y}_{h,il} - M_{h,i} \tilde{r}_{h,il}) \quad (2)$$

and

$$\hat{V}_{h,ij} = \frac{1}{R_{h,i} R_{h,j}} \sum_{k,l \in S_h} \tilde{\Delta}_{h,kl} (\tilde{y}_{h,ik} - M_{h,i} \tilde{r}_{h,ik}) \times (\tilde{y}_{h,jl} - M_{h,j} \tilde{r}_{h,jl}). \quad (3)$$

To evaluate (2) and (3), we make a further approximation by substituting $\hat{R}_{h,i} = \sum_{k \in S_h} \tilde{r}_{h,ik}$ and $\hat{M}_{h,i} = \sum_{k \in S_h} \tilde{y}_{h,ik} / (\sum_{k \in S_h} \tilde{r}_{h,ik})$ for $R_{h,i}$ and $M_{h,i}$.

When sampling rates are small, or if we wish to make predictions for a large super-population (*e.g.*, all potential enrollees in a health plan, not just those currently enrolled), $\tilde{\Delta}_{h,kl} = 1 - \pi_{h,k} \approx 1$ if $k = l$, $\tilde{\Delta}_{h,kl} \approx 0$ if $k \neq l$, and the sampling design approaches sampling with replacement. Under the sampling with replacement design, approximately unbiased estimators are

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{h,i}^2} \sum_{k \in S_h} (\tilde{y}_{h,ik} - \hat{M}_{h,i} \tilde{r}_{h,ik})^2 \quad (4)$$

and

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{h,i} \hat{R}_{h,j}} \sum_{k \in S_h} (\tilde{y}_{h,ik} - \hat{M}_{h,i} \tilde{r}_{h,ik})(\tilde{y}_{h,jk} - \hat{M}_{h,j} \tilde{r}_{h,jk}). \quad (5)$$

These estimators can be generalized to accommodate clustering.

With equal-probability sampling within domains, (4) and (5) reduce to

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{S_{h,i}}^2} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i} r_{h,ik})^2 \quad (6)$$

and

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{S_{h,i}} \hat{R}_{S_{h,j}}} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i} r_{h,ik})(y_{h,jk} - \hat{M}_{h,j} r_{h,jk}), \quad (7)$$

where $\hat{R}_{S_{h,i}}$ is the number of respondents to item i in domain h .

3. Models for Variance Functions

In this section we propose specifications for models for variances and for sample correlations with complete responses or with structured skipped responses. We then discuss model fitting and evaluation strategies. We assume that these domains are nonoverlapping strata, so the sampling errors for different domains are independent.

We transform the ordinal ratings to the $[0, 1]$ interval by the transformation $p_{h,i} = (B_{h,i} - M_{h,i}) / (B_{h,i} - A_{h,i})$, where $A_{h,i}$ and $B_{h,i}$ are the minimum and maximum response categories for item i in domain h respectively. We focus on modeling variances for large values of $M_{h,i}$ (small values of $p_{h,i}$) because in our motivating example mean outcomes are typically near the high end of the scale.

3.1 Variance Functions

To account for the variable number of respondents over domains and items, and differing scales, we normalize the variance estimators in (6) for sample size and re-scale:

$$\tilde{V}_{h,ii} = \frac{\hat{R}_{S_{h,i}} \hat{V}_{h,ii}}{(B_{h,i} - A_{h,i})^2}.$$

With unequal probability sampling within domains, a normalization factor could be used that accounts for the weights. One possible normalization is to multiply $\hat{V}_{h,ii}$ by $\hat{R}_{S_{h,i}}^* = (\sum \tilde{r}_{h,ik})^2 / (\sum \tilde{r}_{h,ik}^2)$, where $\tilde{r}_{h,ik}$ is the response indicator for item i for the k^{th} subject in the h^{th} domain, in place of $\hat{R}_{S_{h,i}}$. This approximation, proposed in Kish (1965), has a model based justification (Gabler, Haeder and Lahiri 1999). It works well if the sampling probabilities vary modestly in the sample, but can lead to inefficiency if the variation is excessive (Korn and Graubard 1999, page 173; Spencer 2000).

Because the items in our example have ordinal scales, the variance must go to 0 as $p_{h,i} \rightarrow 0$ or $p_{h,i} \rightarrow 1$. An obvious predictor with this property is the variance function of the Bernoulli distribution, $p_{h,i}(1 - p_{h,i})$. This holds exactly for

dichotomous items, and might be a useful approximation for items with three or more categories.

As alternatives to the Bernoulli variance model we considered models with a variety of polynomial and other functions of the means as predictors. Of all the models considered, the quadratic family of models were found to fit as well as any. We focused on the following quadratic models.

$$\text{Model V1: } \tilde{V}_{h,ii} = \beta_{1i} p_{h,i}, \quad (8)$$

$$\text{Model V2: } \tilde{V}_{h,ii} = \beta_{2i} p_{h,i} (1 - p_{h,i}), \quad (9)$$

$$\text{Model V3: } \tilde{V}_{h,ii} = \beta_{1i} p_{h,i} + \beta_{2i} p_{h,i} (1 - p_{h,i}). \quad (10)$$

Thus we consider a linear variance model V1, a binomial-like model V2, and a general quadratic variance model V3. All models correctly ensure $\tilde{V}_{h,ii} = 0$ when $p_{h,i} = 0$, but only V2 ensures that $\tilde{V}_{h,ii} = 0$ when $p_{h,i} = 1$. The rationale behind V1 is that relationships are often approximately linear over small intervals. Both V1 and V2 are submodels of the two-parameter quadratic V3. We also considered models for $\log(\tilde{V}_{h,ii})$, but these models did not fit as well.

The model V3 is equivalent to the model suggested by Wolter (1985, chapter 5); the equivalence is seen by expressing the right-hand side of V3 in terms of $p_{h,i}$ and $p_{h,i}^2$, and then dividing both sides by $p_{h,i}^2$ to obtain the relative variance. However, parameter estimates obtained by fitting the two forms of the model may be different depending on the modeling assumptions used.

3.2 Correlation Functions with Complete Data

Because correlations are independent of the scale of the data, we model the correlations and derive the sampling covariances, rather than modeling the covariances directly. We model the sample correlations

$$\hat{\rho}_{h,ij} = \frac{\hat{V}_{h,ij}}{(\hat{V}_{h,ii} \hat{V}_{h,jj})^{1/2}},$$

via the unrestricted transformed values $Z_{h,ij} = \log\{(1 + \hat{\rho}_{h,ij}) / (1 - \hat{\rho}_{h,ij})\}$. Unlike the variance models, models for correlations may include an unrestricted intercept, since there is no natural restriction on the correlation when $p_{h,i}$ or $p_{h,j}$ approaches 0 or 1.

Because $\hat{\rho}_{h,ij}$ is a function of the first and second moments of items i and j , it seemed reasonable to first focus on linear and quadratic models for $Z_{h,ij}$. As with variance functions, we found that a more extensive range of models (e.g., models with logarithms of the means as predictors) did not substantially improve model fit. We ultimately focused on the following nested series of models.

$$\text{Model C1: } Z_{h,ij} = \alpha_{0ij}, \quad (11)$$

$$\text{Model C2: } Z_{h,ij} = \alpha_{0ij} + \alpha_{3ij} p_{h,i} p_{h,j}, \quad (12)$$

$$\text{Model C3: } Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} (p_{h,i} + p_{h,j}) + \alpha_{3ij} p_{h,i} p_{h,j}, \quad (13)$$

$$\text{Model C4: } Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij} p_{h,i} p_{h,j}, \quad (14)$$

$$\text{Model C5: } Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij} p_{h,i} p_{h,j} + \alpha_{4ij} p_{h,i}^2 + \alpha_{5ij} p_{h,j}^2, \quad (15)$$

Model C3 is model C4 with the constraint $\alpha_{1ij} = \alpha_{2ij}$.

3.3 Predicting Covariances with Structured Missing Data

When the data have skip patterns, the sample correlations of the ratings for the set of respondents who answered both items can be modeled by (11)–(15), as in the complete response case. The corresponding sample covariances can be easily estimated by using the fitted variance functions to re-scale the predicted correlations. However, because the sampling covariance reflects the variability in the whole sampling process, not just the variability within the subpopulation of respondents who answered both items, the relationship between sample covariance and sampling covariance is more complicated than if the data were complete. In this section we derive the relationship between the sample covariance for the set of respondents who answered both items and the sampling covariance. This allows correlation models such as (11)–(15) to be applied to data with skip patterns.

There are four distinct data patterns for any pair of items: response to both items, one response and one skipped item (two patterns), and both items skipped. We extend our notation by introducing a superscript representing the response status of a second item. Let $\hat{Y}_{h,ij}^1 = \sum_{S_h} \tilde{y}_{h,ik} \tilde{y}_{h,jk}$, $\hat{Y}_{h,ij}^0 = \sum_{S_h} \tilde{y}_{h,ik} (1 - \tilde{y}_{h,jk})$, $\hat{R}_{h,ij}^1 = \sum_{S_h} \tilde{r}_{h,ik} \tilde{r}_{h,jk}$, $\hat{R}_{h,ij}^0 = \sum_{S_h} \tilde{r}_{h,ik} (1 - \tilde{r}_{h,jk})$, $\hat{M}_{h,ij}^1 = \hat{Y}_{h,ij}^1 / \hat{R}_{h,ij}^1$, $\hat{M}_{h,ij}^0 = \hat{Y}_{h,ij}^0 / \hat{R}_{h,ij}^0$. Then

$$\hat{M}_{h,i} = \frac{\hat{R}_{h,ij}^1 \hat{M}_{h,ij}^1 + \hat{R}_{h,ij}^0 \hat{M}_{h,ij}^0}{\hat{R}_{h,i}}.$$

In the equal probability sampling case, substitution of the above expression for $\hat{M}_{h,i}$ into (7) yields

$$\tilde{V}_{h,ij} = \frac{\hat{R}_{h,ij}^1}{\hat{R}_{h,i} \hat{R}_{h,j}} \left\{ \hat{C}_{h,ij}^1 + \frac{\hat{R}_{h,ij}^0 \hat{D}_{h,ij} \hat{R}_{h,ji}^0 \hat{D}_{h,ji}}{\hat{R}_{h,i} \hat{R}_{h,j}} \right\}, \quad (16)$$

where $\hat{D}_{h,ij} = \hat{M}_{h,ij}^1 - \hat{M}_{h,ij}^0$. Here $\hat{C}_{h,ij}^1 = \sum_S (\tilde{y}_{h,ik} - \hat{M}_{h,ij}^1 \tilde{r}_{h,ik})(\tilde{y}_{h,jk} - \hat{M}_{h,ji}^1 \tilde{r}_{h,jk}) / \hat{R}_{h,ij}^1$ is the normalized sample covariance of the ratings for the set of respondents who answered both items (which can be predicted using correlation and variance functions, and in the case of unequal probability sampling applying a normalization factor). When the sampling probabilities are not equal, Equation (16) holds exactly only if $\sum_S \tilde{r}_{h,jk} (\tilde{y}_{h,ik} - \hat{M}_{h,ik}^1 \tilde{r}_{h,ik}) = 0$. Therefore, (16) may be expected to provide a good approximation if the sampling probabilities for one item are not highly correlated with the residuals for another item. In general, the appropriateness of using (16) for unequal probability sampling designs should be checked.

The estimated mean differences $\hat{D}_{h,ij}$ determine the contribution of the response pattern to the sampling covariance. Either $\hat{D}_{h,ij}$ or $\hat{D}_{h,ji}$ may be modeled in the process of obtaining smoothed estimates of $\tilde{V}_{h,ij}$. In our application, the $\hat{D}_{h,ij}$ were typically small. Because the second term of (16) is a product of two factors of small magnitude ($\hat{D}_{h,ij}$ and $\hat{D}_{h,ji}$), the contribution of $\hat{D}_{h,ij}$ to (16) was small and it sufficed to use a simple model for $\hat{D}_{h,ij}$, such as a constant for each item pair. However, unique constants should be estimated for each pair of items.

3.4 Model Fitting and Evaluation

We estimate the parameters of the variance or correlation function using iteratively reweighted least squares regression. Weighting is important when the number of responses varies greatly across domains, as in our motivating example.

In this section we index domains (h) and respondents (k) but not items as the same methodology applies to each variance and correlation model. Exact computations are derived for the equal probability sampling case, and approximations are noted for the unequal probability sampling case. Generically, the direct estimators \tilde{f}_h , true values f_h , and model predictions \hat{f}_h are related through the hierarchical model

$$\text{Level I: } \tilde{f}_h = f_h + \epsilon_h, \quad (17)$$

$$\text{Level II: } f_h = \hat{f}_h + e_h, \quad (18)$$

where $\epsilon_h \sim [0, \sigma_\epsilon^2 / \hat{R}_{S_h}^*]$, $e_h \sim [0, \tau^2]$, and $[\mu, \sigma^2]$ indicates a distribution with expectation μ and variance σ^2 but unspecified form. In the unequal probability sampling case we replace \hat{R}_{S_h} with $\hat{R}_{S_h}^*$. Here ϵ_h represents sampling error and e_h represents model error. Marginally, $\tilde{f}_h = \hat{f}_h + \epsilon_h + e_h$ so in the regression we weight the observation for domain h by $w_h = (\tau^2 + \sigma_\epsilon^2 / \hat{R}_{S_h}^*)^{-1}$, the inverse of the marginal variance. With equal-probability sampling, the

variance of the direct estimate of $\sigma_h^2 = E[\tilde{f}_h - f_h]^2$ is given by

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{1}{\hat{R}_{S_h} - 1} \left\{ \frac{1}{\hat{R}_{S_h}} \sum_{k \in S} (y_{h,k} - \hat{M}_h r_{h,k})^4 - (1 - \frac{3}{\hat{R}_{S_h}}) \tilde{f}_h^2 \right\} \quad (19)$$

if f is a variance

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{4}{\hat{R}_{S_h} - 3} \text{ if } f \text{ is a transformed correlation.} \quad (20)$$

In the equal probability sampling case Equation (19) is exact and does not depend on parametric assumptions (Seber 1977, page 14). The asymptotic approximation (20) to the variance of the transformed correlation Z_h (Freund and Walpole 1987, page 477) deteriorates as sample sizes decrease, and fails altogether for $\hat{R}_{S_h} \leq 3$. However, domains with small sample sizes have little impact on the fitted models; we exclude domains with $\hat{R}_{S_h} \leq 3$ from correlation modeling.

When the sampling probabilities are not equal, the large sample counterpart to (19), given by

$$\hat{\sigma}_h^2(\tilde{f}_h) = \sum_{k \in S} \left\{ \frac{(\tilde{y}_{h,k} - \hat{M}_h \tilde{r}_{h,k})^2}{\sum_{l \in S} \tilde{r}_{h,l}^2} - \frac{2w_h}{\sum_{l \in S} \tilde{r}_{h,l}} \right\}^2 \times \left(\tilde{y}_{h,k} - \hat{M}_h \tilde{r}_{h,k} - \frac{\tilde{f}_h}{\sum_{l \in S} \tilde{r}_{h,l}^2} \tilde{r}_{h,k}^2 \right)^2,$$

where $w_h = (\sum_S \tilde{y}_{h,l} \tilde{r}_{h,l} / \sum_S \tilde{r}_{h,l}^2) - \hat{M}_h$, may be used. In the equal probability sampling case, $w_h = 0$ and the above expression reduces to a non-bias corrected version of (19). If the sampling probabilities are not equal, we suggest replacing (20) with the design-effect-corrected estimator

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{4}{\hat{R}_{S_h}^* - 3}.$$

The model error variance τ^2 is estimated as:

$$\hat{\tau}^2 = \max \left\{ 0, \hat{MSE} - \frac{\sum \hat{R}_{S_h} \hat{\sigma}_h^2(\tilde{f}_h)}{\sum \hat{R}_{S_h}} \right\},$$

where $\hat{MSE} = \sum_h q_h (\tilde{V}_h - \hat{f}_h)^2$, $q_h = N \hat{R}_{S_h} / \sum_h \hat{R}_{S_h}$, and $N = \sum_h I(\hat{R}_{S_h} > 0)$. The weights are then re-estimated as $\hat{w}_h = (\tau^2 + \hat{\sigma}_h^2(\tilde{f}_h) / \hat{R}_{S_h}^*)^{-1}$, and the GVCf models are refit, iterating to convergence. We again suggest replacing \hat{R}_{S_h} with $\hat{R}_{S_h}^*$ if the sampling probabilities are not equal.

We compared the predictive accuracy of models using $R^2 = 1 - \hat{MSE} / \hat{MSV}$, where \hat{MSE} is the mean squared error of the regression, and \hat{MSV} is the sample size weighted average of the sampling variances of the direct estimators (variances or transformed correlations) for each

domain. Note that we could have $R^2 < 0$ for a very poorly fitting model.

3.5 Combined Estimators

For domains with small samples, direct survey variance estimates often are too imprecise to be useful, while estimates for larger domains in the same study may be quite reliable. Fay and Herriot (1979) and Ghosh and Rao (1994) demonstrated that shrinking direct estimates towards a model-based smoothed value can lead to substantial gains in precision. They proposed composite or empirical Bayes estimators that are weighted averages of direct and model-based estimators. That is, instead of either using the direct estimates or estimates obtained from generalized variance/covariance modeling, we use a weighted average of the two estimators to potentially obtain even better estimates.

Such weighted estimators can be constructed for domain variances using the model specified in (17) and (18). A natural approach is to weight the direct model-based estimators inversely proportional to the corresponding sampling and model error variances respectively (denoted σ_h^2 and τ^2 respectively for domain h). The resulting estimator for domain h (for variances and transformed correlations) is:

$$\tilde{f}_h = \frac{\hat{\tau}^2 \tilde{f}_h^{\text{dir}} + \hat{\sigma}_h^2 \tilde{f}_h^{\text{mod}}}{\hat{\tau}^2 + \hat{\sigma}_h^2} = \tilde{f}_h^{\text{dir}} + \frac{\hat{\sigma}_h^2}{\hat{\tau}^2 + \hat{\sigma}_h^2} (\tilde{f}_h^{\text{mod}} - \tilde{f}_h^{\text{dir}}),$$

where \tilde{f}_h^{dir} and \tilde{f}_h^{mod} denote the direct and model-based estimators. This generic formula applies to the variance estimates for all items, and correlation estimates for all pairs of items. The right-most expression has the form of an empirical Bayes estimator.

If the direct and model-based variance estimators are independent, the variance of the resulting combined estimator is $\tau^2 \sigma_h^2 / (\tau^2 + \sigma_h^2) \leq \min\{\tau^2, \sigma_h^2\}$. Thus the composite is as least as precise as either of its two component estimators, improving on ad hoc selection between direct and model-based predictions. This is a useful strategy especially when model-based predictions improve on direct estimates for some, but not all domains.

4. Example: CAHPS® Data Set

The Consumer Assessments of Health Plans Study (CAHPS®) survey (Goldstein, Cleary, Langwell, Zaslavsky and Heller 2001) was designed primarily to elicit consumer ratings and reports on health plans. Plan mean scores (perhaps after recoding) on the various survey items are calculated and reported to consumers, health plans, and purchasers. Each analytic domain consists of the enrollees of a health plan (or geographically defined portion of one)

in a year; most of the plans are sampled in multiple years. The stratum is the reporting unit (plan or portion thereof) in a given year; reporting units corresponded to plans with the exception of a few large plans that had multiple reporting units. Therefore, there are many units for variance and covariance function estimation.

We illustrate our methods with a CAHPS data set for beneficiaries of U.S. Medicare managed care plans, a system of private but government-funded entities serving from 5.7 to 6.9 million elderly or disabled beneficiaries in each year during our study period (1997 to 2001). Our data represent 381 reporting domains each sampled in 1 to 5 years for a total of 932 distinct reporting unit by year domains with 705,848 responses. Because samples are drawn independently each year, patients may be sampled in multiple years. However, repeated sampling is rare and can be overlooked for our analysis. Therefore, the domains are strata with equal probability element sampling performed within each. Note that in CAHPS analyses no corrections are made for finite-population sampling since the data are collected to guide choices for future years rather than to record experiences of the specific population in a particular year.

CAHPS items use a variety of ordinal response formats with either 11, 4, 3, or 2 response options. Overall ratings of doctor, specialist, care, and plan are measured on a 0 to 10 scale from "worst possible" to "best possible". Other items use a 4-point ordinal "frequency" scale (never/sometimes/usually/always), or a 3-point ordinal "problem" scale (not a problem/somewhat a problem/a big problem), or are dichotomous (no/yes). Many items are answered only by respondents who used particular services or had particular needs, as determined by screener items. For example, an item about whether advice was obtained successfully by telephone is only answered by those who first reported that they attempted to obtain advice in that way.

4.1 Descriptive Statistics

Table 1 presents response distributions and domain mean distributions by item type. Missing observations due to structured skip patterns often occurred in blocks, with as many as 11 items skipped on the basis of a single screening question. Very little nonresponse (less than 2% on almost all items) was not due to a structured skip pattern. In this analysis we treat all types of nonresponse identically.

Item response rates were lowest (as low as 4%) for problem items, several of which dealt with specialty services such as therapy or home health care needed by relatively few respondents. Some of the frequency and yes/no items also had low response rates. The greatest variation in the proportions of skipped items was evident among the yes/no items: 96.7% for a "complaint or problem

with plan” to 12.5% for “get prescription through plan”. Domain mean outcomes are in general concentrated towards the higher end of their scales, indicating that most responses were favorable.

Table 1

Distribution of Responses and Ratings Evaluated over Items of the Same Type (n = 705,848 Respondents)

Statistic	Numerical	How Often	Problem	Yes/No
Number of items	4	11	11	9
Percentage responding				
Mean	74.97	62.56	30.32	57.26
Minimum	50.90	27.70	4.00	12.50
Maximum	95.00	74.50	64.40	96.70
Item means				
Mean	8.76	3.57	2.70	1.78
Minimum	8.57	3.09	2.49	1.62
Maximum	8.88	3.84	2.86	1.97
Distribution of ratings (across items in group)				
0	0.5			
1	0.4	2.0	5.7	19.5
2	0.4	6.3	12.1	80.5
3	0.7	23.9	82.2	
4	0.9	67.8		
5	4.6			
6	3.0			
7	6.2			
8	16.1			
9	17.8			
10	49.5			

Items are on a 0–10 numerical scale from “worse possible” to “best possible”, a 4–point 1–4 ordinal “frequency” scale (never/sometimes/usually/always), a 3–point 1–3 ordinal “problem” scale (not a problem/somewhat a problem/a big problem), or are dichotomous 1–2 items (no/yes).

The domain mean, minimum, and maximum values across all items of the same type are also presented in Table 1. These illustrate that the 0–10 items have the smallest total variation (after rescaling to the common 0–1 range), while the 1–2 items have the largest total variation across domains and items. This is also illustrated in Figure 1, where we observe that the distribution of the 1–2 items varies substantially across items whereas the distributions of the 0–10 items are more homogeneous.

Table 2 presents statistical summary measures for the means and standard deviations of the domain mean ratings, evaluated across items of the same type. This complements Figure 1 by summarizing the difference in distributions of items within a given scale. Items with more response categories are concentrated towards the top of the scale and hence have smaller variance. For example, the mean standard deviation of the 1–2 items (0.36) is twice that of the rescaled 0–10 items (0.172). With the exception of the 0–10 items, the distributions of domain mean ratings vary greatly across items of the same type. For instance, the standard deviation of the means of 1–2 items across items is 0.30 compared to a rescaled standard deviation of 0.03 for the 0–10 items.

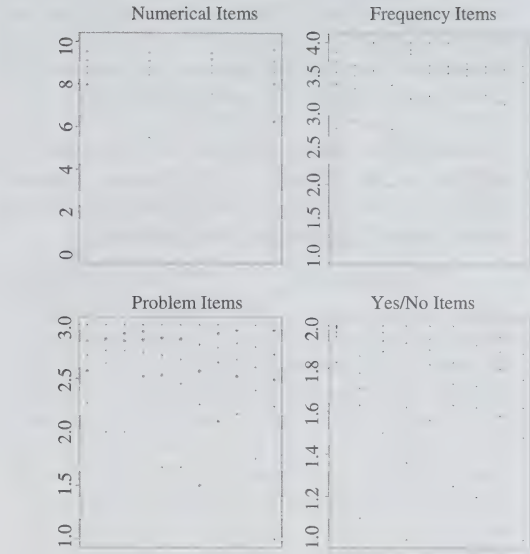


Figure 1. Five-point Summary of the Domain Sample Means for Each Item. The five-point summary consists of the minimum, 10th percentile, mean, 90th percentile, and the maximum.

Table 2
Summary Statistics of Domain Means and Standard Deviations Evaluated Over Domains and Items

Type	Summary Statistics for:					
	Item Means			Item SDs		
	Min	Max	Mean	SD	Mean	SD
Numerical 0–10	6.82	9.52	8.76	0.30	1.72	0.26
Frequency 1–4	2.86	3.90	3.57	0.12	0.66	0.09
Problem 1–3	1.88	2.99	2.70	0.14	0.57	0.13
Yes/No 1–2	1.34	1.96	1.78	0.08	0.36	0.06

Note: Columns 2 through 5 give the minimum, maximum, mean, and standard deviation of the domain item means across items of a given type. Columns 6 and 7 give the mean and standard deviation of the domain item standard deviations across items of a given type.

Sample correlations also varied greatly across the pairs of items (Figure 2), although most were positive. Correlations between items of the same type most often were higher than those between items of different types. The numerical 0–10 ratings had the largest correlations (mean = 0.49), and generally ratings with more categories tended to have higher correlations than ratings with fewer categories. Although most of the pairs of 1–4 items had mean correlations near to 0.5, one item was negatively correlated with the others (revealed by the cluster of mean correlations below 0); this arose from reverse coding an item whose overall sample mean was not in the top half of the scale. The distributions

of the correlations of pairs of 1–2 items were centered near 0, indicating that pairs of items of this type often have negative correlations. Complete item wordings and additional summary statistics appear in Zaslavsky, Beaulieu, Landon and Cleary (2000) and Zaslavsky and Cleary (2002).

Models fitted to the variances and correlations are presented in the remainder of this section. Extensive checking of the best-fitting models indicated that the residuals did not follow any discernible pattern.

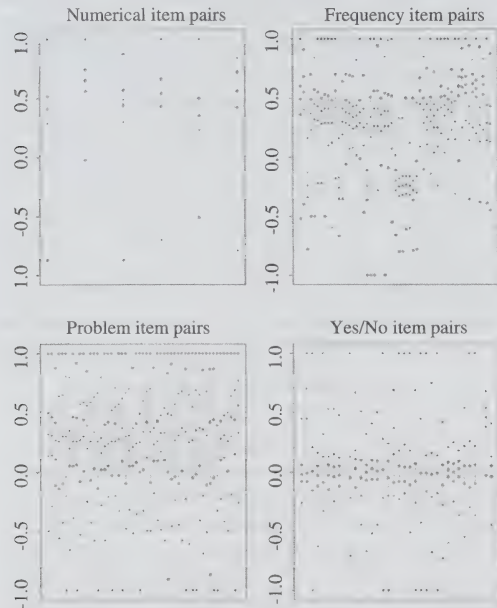


Figure 2. Five-point summary of the domain sample correlations between items with the same type. The five-point summary consists of the minimum, 10th percentile, mean, 90th percentile, and the maximum.

4.2 Variance Functions

In preliminary investigations not reported here, we fit two models within groups of items with the same response scale, one with common and one with different regression parameters for each item, to the data set comprising all of the items. Comparisons of the overall fits of the models (using criteria such as Mallow’s C_p , R^2 , adjusted R^2) and tests of the significance of effect-item interactions demonstrated that allowing parameters to vary across items significantly improved model fit. For instance, for the rescaled numerical ratings, weighted by domain sample size, the two models’ root mean squared errors were 0.446 versus 0.402, and values of R^2 were 0.783 versus 0.825. Based on this we decided to fit separate models for each item.

The variance functions (8–10) were fitted to each item except the yes/no items, which follow the binomial variance function in the equal-probability sampling case. The iterative procedure described in section 3.4 converged almost precisely in exactly two iterations. This is because the weights for the observations change only with the estimate of τ^2 , and so very little change in the weights occurs after the first iteration.

Table 3 presents the average sampling variation, average model error variation, and R^2 , for each model averaged over items of each response scale. Sampling variation, computed using (19), does not depend on the model.

Table 3 Goodness-of-fit Statistics for Variance Functions						
Rating Scale	0–10		1–4		1–3	
Sampling Variation	0.1460		0.3511		3.1703	
	ModErr	R^2	ModErr	R^2	ModErr	R^2
Model V1	0.020	0.741	0.066	0.824	0.069	0.916
Model V2	0.043	0.710	0.036	0.835	0.000	0.940
Model V3	0.016	0.750	0.024	0.847	0.000	0.947
Prob(ModErr < Sampling Variation)						
Model V1	0.968		0.916		0.996	
Model V2	0.858		0.967		0.996	
Model V3	0.981		0.983		0.996	

ModErr is the variance component for lack of fit, R^2 is as defined in section 3.4, Prob(ModErr < Sampling Variation) is the proportion of domains for which model error is smaller than sampling variation. All ratings are rescaled to a 0–1 scale, and model errors are multiplied by 10⁴.

For items with few categories (more closely resembling the binomial), the quadratic component of the variance function tends to dominate the linear component, making models V2 and V3 fit better than V1. Because V2 imposes a constraint at a point far outside the range of the domain means, it does not fit the data as well when there are more categories and the data are consequently further from binomial. The 0–10 items are less dispersed than the 1–4 and 1–3 ratings, enabling the linear model to fit better. The R^2 values for model V3 were close to 0.75 for numerical (0–10) items, 0.85 for the frequency (1–4) items, and 0.95 for the problem (1–3) items.

The lower portion of Table 3 displays for each item the proportion of domains (of those with at least 2 responses to the given item) for which sampling variation is larger than model error variation. For over 90% of domains, model error variation was less than the sampling variation of the direct variance estimate.

Figure 3 illustrates the fit of V3 for two each of the 0–10, 1–4, and 1–3 items. Illustrations for the remaining items are similar, but are not provided due to space limitations. The fitted curves are constrained to 0 at the maximum ratings. To assess the impact this constraint has on the fitted

variance function, we also fit an unrestricted (three parameter) quadratic variance function; these attained values very close to 0 at the maximum rating, and closely approximated the fitted curve from the constrained models, further supporting V3.

Average parameter estimates and their standard deviations over items of the same type are shown in Table 4. The parameters differed substantially across items, supporting the decision to estimate separate regression coefficients. In most cases the coefficients for both the $p_{h,i}$ and $p_{h,i}(1-p_{h,i})$ terms in V3 were significant, indicating that these are needed for generalized variance modeling. In some cases (particularly with the 0–10 items) the coefficient of the $p_{h,i}(1-p_{h,i})$ term was negative, resulting in an estimated variance function that is convex rather than concave (the shape of the binomial variance function). This can happen when the sample means for the ratings are concentrated on a small proportion of the response scale, over which the linear term explains much of the variation in the data. As mentioned earlier, adding higher-order polynomial or logarithmic functions of $p_{h,i}$ did not significantly improve model fit.

Table 4
Average Variance Function Parameter Estimates for Each Type of Item and Standard Deviations Across Items (in Parentheses)

Model	Item Type					
	0–10		1–4		1–3	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
V1	0.236 (0.016)	–	0.354 (0.039)	–	0.569 (0.068)	–
V2	–	0.271 (0.020)	–	0.421 (0.034)	–	0.711 (0.069)
V3	0.334 (0.143)	–0.114 (0.155)	0.151 (0.104)	0.241 (0.132)	0.239 (0.112)	0.420 (0.110)

See Table 1 for a description of the 0–10, 1–4, and 1–3 items.

4.3 Correlation Functions

Models are ordered from simplest (C1, the constant model) to most complex (C5, containing all linear and quadratic terms). As for the variance models, statistical tests found highly significant item interaction effects, implying that separate models should be fit for each pair. We did not expect all pairs of items to have similar correlations, since by intention the items are divided into internally consistent groups, each of which measures a distinct aspect of patient experiences such as interactions with doctor or dealings with customer service agents (Hays, Shaul, Williams, Lubalin, Harris-Kojetin, Sweeny and Cleary 1999).

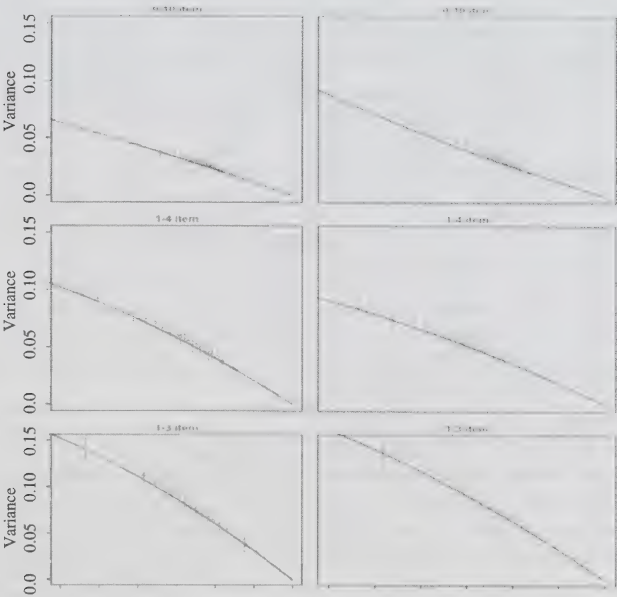


Figure 3. Quadratic Variance Function (V3) of Two Items for each Rating Type. Each point is the average of 60 domains. Vertical lines join the 10th and 90th percentiles of the distribution of the variances. For this and following displays the direction of the transformed horizontal axis has been reversed to agree with that of the original variables.

The fits of the correlation models for pairs of items of the same type are summarized in Table 5. Over the range of models considered, the biggest improvements in model performance (as measured by R^2) occur between model C1 and model C2, and between model C3 and model C4. For example, the average R^2 for the numerical ratings in models C3–C5 are 0.0391, 0.1494, and 0.1508 respectively, and the average R^2 for the 1–4 ratings over C1–C3 are 0, 0.0700, and 0.0789 respectively. This suggests that C2 and C4 are the best models for different pairs of items, a claim that is supported by the hypothesis tests on the significance of the incremental improvements in model fit.

Sampling variation was highest for the 1–3 ratings, at least in part because high rates of non-response due to skipped responses diminished the sample sizes. Model error and R^2 of correlation models for items of different types were similar to those for models for items having the same type.

The R^2 values for the correlation models were between 0.029 and 0.15 for all pairs of items. Although there was no evidence to suggest that C4 was an inappropriate model for the correlations, these results indicate that substantial variation in the correlations is not explained by the item means.

The sampling variances of the direct estimates were often less than the corresponding model error variances (lower part of Tables 5 and 6 especially for the 0–10 items. Under C4, model error variances were smaller for only 13% of domains for the 0–10 ratings, 45% of domains for the 1–4 ratings, and approximately 81% of domains for the 1–3 and 1–2 ratings.

Figure 4 presents the observed correlations and fitted function C4 for an illustrative pair of items from each of the 10 combinations of item types, representing the 595 distinct pairs of items. To illustrate the fitted correlation models, we adjust the observed and fitted correlations to the mean of one item and plot the resulting values in two-dimensional space. This process is repeated for the other item, yielding two plots for each correlation.

Figure 4 illustrates the generally weak relationship of the correlation to the means of the items seen in Tables 5 and 6. Analysis of Tables 5 and 6 reveals that the relationship between the correlation and the mean outcome is weaker for items with fewer categories and with correlations of items of different types. In particular, the 0–10 numerical ratings are the only group for which there is a clear correlation-mean relationship.

Table 5

Model Fitting Diagnostics for Correlation Functions for Items of the Same Type, Averaged over Pairs of Items of the Same Type

Rating Type	0 – 10		1 – 4		1 – 3		1 – 2	
Sampling Variation	0.0124		0.0178		0.1482		0.0325	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
Model C1	0.060	0.000	0.028	0.000	0.112	0.000	0.018	0.000
Model C2	0.060	0.013	0.025	0.070	0.103	0.048	0.017	0.014
Model C3	0.057	0.039	0.024	0.079	0.102	0.054	0.017	0.018
Model C4	0.047	0.150	0.023	0.100	0.100	0.068	0.016	0.029
Model C5	0.044	0.151	0.023	0.105	0.096	0.080	0.015	0.034
Prob(ModErr < Sampling Variation)								
Model C1	0.033		0.339		0.461		0.788	
Model C2	0.033		0.400		0.498		0.795	
Model C3	0.034		0.411		0.502		0.796	
Model C4	0.038		0.435		0.516		0.799	
Model C5	0.065		0.440		0.530		0.802	

See Table 1 for a description of the 0–10, 1–4, 1–3 and 1–2 items, and Table 3 for an explanation of the column headings.

Table 6

Model Fitting Diagnostics for Correlation Functions for C4 by Type of Item.
Averaged over Items of the Same Type

Types	0 – 10		1 – 4		1 – 3		1 – 2	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
0–10	0.047	0.149	0.021	0.104	0.040	0.094	0.013	0.059
1–4			0.023	0.100	0.038	0.076	0.013	0.039
1–3					0.100	0.068	0.028	0.031
1–2							0.016	0.029
Prob(ModErr < Sampling Variation)								
0–10	0.038		0.358		0.523		0.784	
1–4			0.435		0.605		0.790	
1–3					0.516		0.827	
1–2							0.799	

See Table 1 for a description of the 0–10, 1–4, 1–3 and 1–2 items, and Table 3 for an explanation of the column headings.

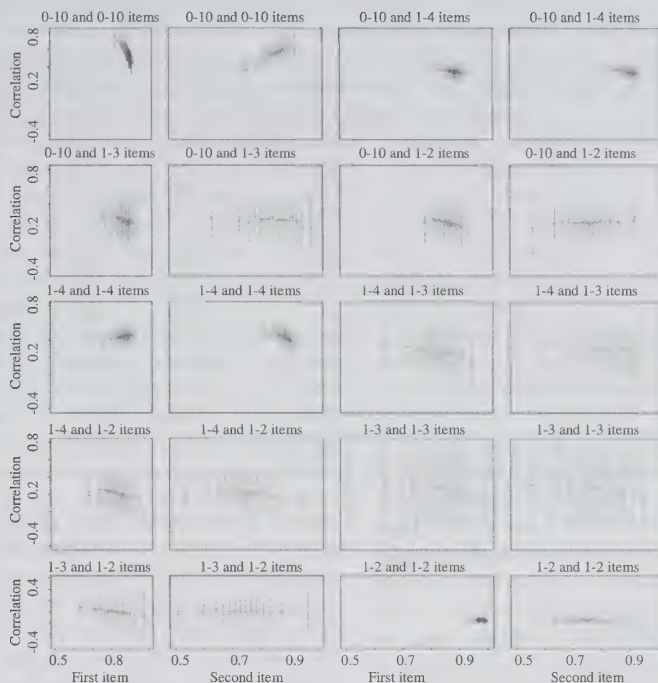


Figure 4. Correlation Functions for One Pair of Items for Each Combination of Rating Types.

Note: The plots for each items involved in the correlation are side by side. Refer to Figure 3 for a description of the contents and axes of the plot.

Although the fitted curves for the correlation functions are nearly flat, the variation in the parameter estimates under model C4 for α_4 are large and were suggestive of instability. The wildly varying parameter estimates are a consequence of collinearity among the predictors in model C4. In many cases the estimated value of α_4 offsets the parameter estimates for the linear predictors, resulting in a fitted curve that is nearly flat.

4.4 Mean Difference Functions

The difference $\hat{D}_{h,ij}$ appeared to depend on both the marginal mean and its square, implying a model analogous to V3 could be appropriate. However, because $\hat{D}_{h,ij}$ typically is small enough that $\hat{D}_{h,ij} \hat{D}_{h,ji}$ has minimal impact on (16), we fit a constant model.

4.5 Composite Estimator

Table 7 presents the quantiles of the distribution of weights $\sigma_h^2 / (\tau^2 + \sigma_h^2)$ for the model-based estimate, used in the composite estimator of section 3.5, averaged over items (or pairs of items) of the same type. The proportion of

domains for which the standard error of the model-based predictions was smaller than that of the direct estimates is also presented. As noted previously, the model-based predictions have more weight in the composite variance estimates than in the composite correlation estimates. The average (across items or pairs) median of the weights of the model-based estimator ranged from 0.892 to 1.000 for variances, 0.256 to 0.709 for correlations of items of the same type, and from 0.468 to 0.738 for correlations of items of different types. Also, for both variances and correlations, the weight of the model-based predictions was larger for items with fewer response categories. For example, the model-based estimator had median weights of 0.256, 0.468, 0.540, and 0.647 on the composite estimates of correlations when the numerical 0–10 ratings were paired with the 0–10, 1–4, 1–3, and 1–2 ratings, respectively. However, even for pairs of 0–10 numerical ratings, for which sampling error of the direct estimator exceeded the model error in only 3.81% of domains, these results indicate that the median weight of the model-based estimator was 0.256, a nontrivial amount.

Table 7
Distribution of Weights for the Model-Based Component of the Composite Estimator, Averaged Over Items of Same Type

Model	Item Type		Prob(ModErr < Sampling Variation)	Quantiles		
	1	2		10%	Median	90%
Variance	0–10	–	0.981	0.778	0.892	0.948
	1–4	–	0.983	0.948	0.966	0.974
	1–3	–	0.996	1.000	1.000	1.000
Correlation	0–10	0–10	0.038	0.141	0.256	0.335
	0–10	1–4	0.358	0.301	0.468	0.562
	0–10	1–3	0.523	0.357	0.540	0.654
	0–10	1–2	0.784	0.531	0.695	0.767
	1–4	1–4	0.435	0.324	0.497	0.591
	1–4	1–3	0.605	0.404	0.587	0.699
	1–4	1–2	0.853	0.584	0.738	0.805
	1–3	1–3	0.516	0.349	0.540	0.675
	1–3	1–2	0.827	0.584	0.737	0.817
	1–2	1–2	0.799	0.541	0.709	0.780

The distribution of weights is summarized by the 10th, 50th, and 90th percentiles. See Table 3 for definition of ModErr.

4.6 Joint Predictions

Because we modeled the correlations independently for each item, our fitted correlation matrices do not necessarily satisfy the constraint of positive definiteness, which can be important for multivariate inference. In additional work, we have determined that as long as the multivariate analysis is restricted to items of the same type, the fitted correlations from the C2 and C4 models yield positive definite estimates of correlation matrices for almost all domains. However, for analyses including items of different types (*e.g.*, the 0–10 numerical items, and the 1–2 yes/no items), predictions based on C4 predict correlation matrices that are indefinite for many domains, while predictions based on C2 are more stable and almost always yield positive definite predictions. This suggests that while C4 may be slightly superior in terms of univariate model fit, C2 may be more appropriate for multivariate inference.

One way of overcoming the problem of indefinite predicted correlation matrices is to use a weighted average of the predicted correlation matrix for a domain and the estimated average correlation matrix (EACM) across domains. The EACM may be constructed by weighting the direct estimates (each of which is at least positive semi-definite) by the total sample size for each domain. Then any indefinite predicted correlation matrices are replaced with the weighted average of the predicted correlation matrix and the EACM, where the weight used for each domain is increased until a positive definite matrix results. Like an empirical Bayes estimator, this process stabilizes estimates by effectively shrinking the model coefficients toward those of a simpler (constant) model.

When analyzing all 35 CAHPS items simultaneously the EACM had an average weight across domains of 0.65 with

model C4, whereas with model C2 the average weight was only 0.01 since the predicted correlations under C2 were usually positive definite. In analyzing only the 0–10, 1–4, and 1–3 items the EACM had average weights of 0.28 and 0.00 with C4 and C2 respectively, while in analyzing just the 0–10 and 1–4 items the corresponding average weights were 0.06 and 0.00. When analyzing the different types of items separately, the average weight of the EACM with C4 was 0.00 for the 0–10 and 1–4 items, 0.01 for the 1–3 items, and 0.17 for the 1–2 items. The EACM is thus not needed when analyzing the 0–10 and 1–4 items because the predicted correlation matrices were positive definite for every domain.

5. Conclusion

We have presented methodology for estimating variance and covariance functions for domain means of ordinal survey items. Our methodology can also be applied to survey items measured on continuous scales. We introduced a decomposition of the model error that allows the variation due to sampling to be separated from that due to model fit. The decomposition also helps to avoid over-fitting because it estimates the proportion of variation in the data that can be modeled and thus when the current predictors suffice.

The procedure for fitting the variance and correlation models is the same regardless of whether or not the data contain skip patterns. The analytic derivation in section 3.3 shows that if skip patterns are present, mean differences of items by response status of other items are required in order to compute the sampling covariance estimates. However, we argued that these quantities are likely to have minimal impact on the results and that therefore a constant model

could be used, which was supported by our empirical findings.

A quadratic variance function constrained to 0 at the maximum rating, and a model for transformed correlations involving the product but not the squares of the means, best predicted the direct estimates in our applied example. The modeled variance estimates generally had much smaller standard errors than the direct estimates; the same was, however, not true of the correlation estimates. It is interesting and reassuring that our quadratic variance function can be expressed as the widely-used relative variance model of Wolter (1985).

For our ordinal data, the estimates of the domain mean ratings contain minimal information about the correlation between the ratings. Hence, the mean-covariance relationship is principally an artifact of the mean-variance relationship. However, for items with many response categories, the association between correlations and mean outcomes for items of the same type was stronger most notably for pairs of 0–10 items. With the exception of the 0–10 and possibly the 1–4 ratings, the correlations might as well be modeled as constants, which also makes it easier to guarantee positive definiteness of the predicted correlation matrix. However, it is important that the parameters of the correlation model be allowed to vary across pairs of items.

A composite estimator that weights the direct and model-based estimators proportional to their precisions has smaller variance than either estimator alone, especially when the components have close to equal weight. The model-based estimator had the greatest influence on estimates for small domains, for which little information is available. The model-based estimator had the greatest influence on estimates for variances, followed by correlations of items of the same type, and lastly correlations of items of different types. Both model-based and composite estimators can be benchmarked (ratio adjusted) to agree on the average across domains with direct estimates, although this proved to be unnecessary in our example.

GVCFs find several applications in our continuing research. We are developing quasi likelihood-based methods for estimating covariance matrices for the domain means of ordinal survey items, representing the second-level (structural) covariance in a hierarchical model (O'Malley and Zaslavsky 2004). GVCF models are needed to provide estimates of sampling variances and covariances and to modify those estimates as the means are re-estimated during the fitting procedure. If the sampling variability of the GVCF estimates is minimal because the number of domains is large, the GVCF predicted variances and covariances can be treated as known. However, if the sampling error of the GVCF-based estimates is large a model that allows these errors to propagate through the analysis should be used. In

related work, Fay and Train (1997) used a binomial model with a design effect for each domain in empirical Bayes estimation of binomial rates. Our research extends this approach to multivariate estimation and more general response formats.

Another application of GVCFs is the computation of variance estimates for linear combinations of item means, facilitating variance estimation for composite scores, like those used in CAHPS reporting. The methods described in section 2 are applicable to variance estimation for any functions of totals, including functions of means, other ratios, or regression coefficients.

There are several ways of extending the GVCF methodology. In addition to summary measures of outcomes, generalized variance and covariance functions (GVCFs) may also depend on other independent variables, in particular those that would better predict correlations. We considered variables summarizing response patterns, such as the proportion of respondents in a domain, but these did not improve the model. GVCFs could also be extended to multi-stage sampling.

Acknowledgements

This work was supported by the U.S. Agency for Healthcare Research and Quality through the Consumer Assessments of Health Plans Study (grant U18 HS09205-06) and by the U.S. Centers for Medicare and Medicaid Services (contract 500-95-007). We thank Paul D. Cleary for his ongoing support of this work, Matt Cioffi for data management, and Elizabeth Goldstein and Amy Heller of the Centers for Medicare and Medicaid Services (CMS), and the other members of the CAHPS-MMC survey implementation team.

References

- Cho, M.J., Eltinge, J.L., Gershunskaya, J. and Huff, L.L. (2002). Evaluation of generalized variance function estimators for the U.S. Current Employment Survey. In *Proceedings of the Joint Statistical Meetings* [CDROM]. Alexandria, VA: American Statistical Association, 534-539.
- Eltinge, J. (2002). Use of generalized variance functions in multivariate analysis. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 904-913.
- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Fay, R.E., and Train, G.F. (1997). Small domain methodology for estimating income and poverty characteristics for states in 1993. In *Proceedings of the Social Statistics Section*, Alexandria, VA: American Statistical Association, 183-188.

- Freund, J.E., and Walpole, R.E. (1987). *Mathematical Statistics*. New Jersey: Prentice-Hall, Inc., 4th Edn.
- Gabler, S., Haeder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.
- Goldstein, E., Cleary, P.D., Langwell, K.M. Zaslavsky, A.M. and Heller, A. (2001). Medicare Managed Care CAHPS: A tool for performance improvement. *Health Care Financing Review*, 22, 101-107.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.
- Hays, R.D., Shaul, J.A., Williams, V.S.L., Lubalin, J.S., Harris-Kojetin, L.D., Sweeny, S.F. and Cleary, P.D. (1999). Psychometric properties of the CAHPS 1.0 survey measures. *Medical Care*, 37 (Supplement), 22-31.
- Huff, L.L., Eltinge, J.L. and Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. Current Employment Survey. In *Proceedings of the Joint Statistical Meetings* [CDROM], Alexandria, VA: American Statistical Association, 1519-1524.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.
- O'Malley, A.J. and Zaslavsky, A.M. (2004). Implementation of cluster-level covariance analysis for survey data with structured nonresponse. In *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 1907-1914.
- Otto, M.C., and Bell, W.R. (1995). Sampling error modeling of poverty and income statistics for states. In *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.
- Spencer, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*, 26, 137-138.
- Valliant, R. (1992a). Longitudinal smoothing of price index variances. In *Statistics Canada Symposium*. Ottawa: Statistics Canada. 113-120.
- Valliant, R. (1992b). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 8, 433-444.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Woodruff, S. (1992). Variance estimation for estimates of employment change in the Current Employment Statistics Survey. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA: American Statistical Association, 626-631.
- Zaslavsky, A.M., Beaulieu, N.D., Landon, B.E. and Cleary, P.D. (2000). Dimensions of consumer-assessed quality of Medicare managed-care health plans. *Medical Care*, 38, 162-174.
- Zaslavsky, A.M., and Cleary, P.D. (2002). Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey. *Medical Care*, 40, 951-964.

Spatio-Temporal Models in Small Area Estimation

Bharat Bhushan Singh, Girja Kant Shukla and Debasis Kundu¹

Abstract

A spatial regression model in a general mixed effects model framework has been proposed for the small area estimation problem. A common autocorrelation parameter across the small areas has resulted in the improvement of the small area estimates. It has been found to be very useful in the cases where there is little improvement in the small area estimates due to the exogenous variables. A second order approximation to the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP) has also been worked out. Using the Kalman filtering approach, a spatial temporal model has been proposed. In this case also, a second order approximation to the MSE of the EBLUP has been obtained. As a case study, the time series monthly per capita consumption expenditure (MPCE) data from the National Sample Survey Organisation (NSSO) of the Ministry of Statistics and Programme Implementation, Government of India, have been used for the validation of the models.

Key Words: Mixed effects linear model; Spatial autocorrelation; Weight matrix; Best linear unbiased predictor; Empirical best linear unbiased predictor; Kalman filtering; NSSO rounds.

1. Introduction

Local level planning requires reliable data at the appropriate level. The complete enumeration or large sample surveys with adequate sample size is expensive and time consuming. The censuses are usually carried out once in a decade, while the sample surveys are often planned to provide estimates at much higher level. One such large sample survey is socio-economic survey of National Sample Survey Organisation (NSSO). Here the direct survey estimates are available at small area (district) level as most of the districts are stratum in the sampling procedure adopted by the NSSO. However, the estimates are exceedingly unreliable due to unacceptably large standard errors. This requires strengthening of such estimates with the use of information from similar small areas or with the help of some reliable exogenous variables, easily available and related to the variable under study.

Various model based approaches have been suggested to improve the direct estimators. The model-based approach facilitates its validation through the sample data. The simple area specific model suggested is two stage model of Fay and Herriot (1979).

$$y_i = \theta_i + \varepsilon_i, \quad E(\varepsilon_i | \theta_i) = 0, \quad \text{Var}(\varepsilon_i | \theta_i) = \sigma_\varepsilon^2, \quad (1.1)$$

$$\theta_i = X_i^T \beta + v_i, \quad E(v_i) = 0, \quad \text{Var}(v_i) = \sigma_v^2, \quad i = 1, 2, \dots, m. \quad (1.2)$$

Here y_i 's are direct survey estimators of θ_i 's, the characteristic under study. θ_i 's may be population small area means. $X_i = (X_{i1}, \dots, X_{ip})^T$'s are exogenous variables which are available and assumed to be closely related to θ_i 's and z_i 's are known positive constants. $\beta(p \times 1)$ is the vector of regression parameters.

The first equation (1.1) is the design model while the second (1.2) is the linking model. The ε_i 's are sampling errors. Estimators y_i 's are design unbiased and the sampling variances σ_ε^2 's are known. Further the ε_i 's and v_i 's are identically and independently distributed random variables. Normality of the random errors and random effects are often assumed. For this model, best linear unbiased predictor (BLUP) on the line of the best linear unbiased estimator (BLUE) has been suggested. The estimate is design consistent and model unbiased (Ghosh and Rao 1994). It is typically the weighted average of the direct survey estimator y_i and the regression synthetic estimator $X_i^T \beta$. The BLUP estimator depends on variance component σ_v^2 which is unknown in practical applications. Various methods of estimating variance components in general mixed effects linear model are available (Cressie 1992). By replacing σ_v^2 with an asymptotically consistent estimator $\hat{\sigma}_v^2$, an empirical best linear unbiased predictor (EBLUP) has also been obtained.

The main problem associated with the data in the Indian context is the non-availability of administrative or civic registration data at small area level. Often, it is difficult to find out the exogenous variables closely related (multiple correlation coefficient $R^2 > 0.5$) to the variable under study.

In the present paper, the exploitation of spatial autocorrelation amongst the small area units in the form of spatial model, has been considered for improving the small area estimators. Besides this, for the time series data, a spatial temporal model on the line of Kalman filtering has been utilised to further improve the estimators. Time series data on monthly per capital consumption expenditure

1. Bharat Bhushan Singh, Girja Kant Shukla and Debasis Kundu, Department of Mathematics, I.I.T. Kanpur-208016. E-mail: drbbsingh@hotmail.com.

(MPCE) as estimated from a large sample survey carried out by the National Sample Survey Organisation (NSSO) has been studied. In the present paper, we propose suitable models in the framework of mixed effects linear model to provide better estimators of the MPCE at small area level.

Rest of the paper has been organized as follows. In Section 2, we consider a Spatial Model on the line of general mixed effects linear model with the introduction of spatial autocorrelation among the small area units. The BLUP and EBLUP of the mixed effects have been presented. A second order approximation to the MSE of the EBLUP and to the estimator of the MSE has also been obtained. Section 3 deals with the time series extension of Spatial Model in form of Spatial Temporal Model, using the Kalman filtering approach. The BLUP and the EBLUP of the mixed effects along with a second order approximation to the MSE of the EBLUP and to the estimator of the MSE have been discussed. Section 4 presents and analyses estimates of the MPCE from a large sample survey carried out periodically in India. The conclusions of the data analysis are reported in Section 5. All the proofs have been provided in the Appendix.

2. Spatial Model

The small area characteristics usually have the spatial dependence in terms of neighbourhood similarities. Cressie (1990) used conditional spatial dependence among random effects, in the context of adjustment for census undercounts. Here, we use simultaneous spatial dependence (Cliff and Ord 1981) among the random effects which has certain advantage over conditional dependence (Ripley 1981). We have thus tried to explain a portion of the random error unaccounted for and left over by explanatory variables which makes it possible to improve the direct survey estimators. The proposed model is a three stage area specific model (Ghosh and Rao 1994).

$$y = \theta + \varepsilon, \quad \varepsilon \sim N_m(0, R), \quad (2.1)$$

$$\theta = X\beta + u, \quad (2.2)$$

$$u = \rho W u + v, \quad v \sim N_m(0, \sigma_v^2 I), \quad (2.3)$$

where θ is a m -component vector (corresponding to number of small areas) for the characteristic under study and y is its direct survey estimator obtained through small sample data. In the above model, the first equation (2.1) shows the design (sampling) model, the second equation (2.2) shows regression model and the third one (2.3) shows spatial model on the residuals, the later two are linked in the first equation. The above model can be expressed as

$$y = X\beta + Zv + \varepsilon, \quad Z = (I - \rho W)^{-1}, \quad (2.4)$$

where $X(m \times p)$ is the design matrix of full column rank p , $\beta(p \times 1)$ is a column vector of regression parameters and $Z(m \times m)$ represents the coefficients of random effects v . $W(m \times m)$ is a known spatial weight matrix which shows the amount of interaction between any pair of small areas. The elements of $W \equiv [W_{ij}]$ with $W_{ii} = 0 \quad \forall i$ may depend on the distance between the centers of small areas or on the length of common boundary between them. As a simple alternative, it may have binary values $W_{ij} = 1$ (unscaled) if j^{th} area is physically contiguous to i^{th} area and $W_{ij} = 0$, otherwise. The matrix has been standardised so as to satisfy $\sum_{j=1}^m W_{ij} = 1$ for $i = 1, 2, \dots, m$. The constant ρ is a measure of the overall level of spatial autocorrelation and its magnitude reflects the suitability of W for given y and X . Further v and ε are assumed to be independent of each other. R is a diagonal matrix of order m which may be expressed as $R = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2)$ where σ_i^2 's are known sampling variances corresponding to the i^{th} area. The parameter vector $\psi = [\rho, \sigma_v^2]^T$ has two elements.

In this model the strength is borrowed from the similar small areas through two common parameters viz. regression parameter β and autocorrelation parameter ρ . Note that the present model is a more general model and the model of Fay and Herriot (1979) can be obtained from this by taking $\rho = 0$.

By adopting the mixed effects linear model approach (Henderson 1975), the best linear unbiased predictor (BLUP) of $\theta = X\beta + Zv$ and the mean squared error (MSE) of the BLUP may be obtained as

$$\begin{aligned} \hat{\theta}(\psi) &= X\hat{\beta}(\psi) + \Lambda(\psi)[y - X\hat{\beta}(\psi)] \\ &= \sigma_v^2 A^{-1}(\psi) \Sigma^{-1}(\psi) y + R \Sigma^{-1}(\psi) X \hat{\beta}(\psi), \end{aligned} \quad (2.5)$$

$$\begin{aligned} \text{MSE}[\hat{\theta}(\psi)] &= \\ E[(\hat{\theta}(\psi) - \theta)(\hat{\theta}(\psi) - \theta)^T] &= g_1(\psi) + g_2(\psi), \end{aligned} \quad (2.6)$$

$$g_1(\psi) = \Lambda(\psi) R = R - R \Sigma^{-1}(\psi) R, \quad (2.7)$$

$$g_2(\psi) = R \Sigma^{-1}(\psi) X (X^T \Sigma^{-1}(\psi) X)^{-1} X^T \Sigma^{-1}(\psi) R, \quad (2.8)$$

$$\hat{\beta}(\psi) = [X^T \Sigma^{-1}(\psi) X]^{-1} X^T \Sigma^{-1}(\psi) y,$$

$$\Sigma(\psi) = \sigma_v^2 A^{-1}(\psi) + R,$$

$$\Lambda(\psi) = \sigma_v^2 A^{-1}(\psi) \Sigma^{-1}(\psi), \quad A(\psi) = (I - \rho W)^T (I - \rho W).$$

Here $\hat{\beta}$, Σ and A , all are the functions of ψ and usually have been expressed as $\hat{\beta}(\psi)$, $\Sigma(\psi)$ and $A(\psi)$ respectively. However, sometimes due to brevity, the suffix ψ has been omitted. The first term, $g_1(\psi)$ in the expression for the MSE, shows the variability of $\hat{\theta}$ when all the parameters are known and is of order $O(1)$. The second term, $g_2(\psi)$, due to estimating the fixed effects β , is of order $O(m^{-1})$ for large m . Further, with $\rho = 0$, the above

model reduces to the standard mixed effects linear regression model while for $X\beta = \mu$, we obtain a purely spatial scheme with only intercept term.

In practice parameter ψ is unknown and is estimated from the data. The maximum likelihood estimator (MLE) of the parameter, ψ is obtained by maximizing the following log likelihood function of ψ

$$l = \text{const} - \frac{1}{2} \log [|\Sigma(\psi)|] - \frac{1}{2} [y - X\hat{\beta}(\psi)]^T \Sigma^{-1}(\psi) [y - X\hat{\beta}(\psi)] \quad (2.9)$$

with respect to the parameter ψ . The empirical best linear unbiased predictor (EBLUP), $\hat{\theta}(\psi)$ and the naive estimator of the MSE are obtained from the equations (2.5) and (2.6) respectively, by replacing the parameter vector ψ by its estimator $\hat{\psi}$.

$$\hat{\theta}(\hat{\psi}) = \hat{\sigma}_v^2 A^{-1}(\hat{\psi}) \Sigma^{-1}(\hat{\psi}) y + R \Sigma^{-1}(\hat{\psi}) X \hat{\beta}(\hat{\psi}), \quad (2.10)$$

$$\text{MSE}[\hat{\theta}(\hat{\psi})] = g_1(\hat{\psi}) + g_2(\hat{\psi}), \quad (2.11)$$

$$\text{where } \Sigma(\hat{\psi}) = \hat{\sigma}_v^2 A^{-1}(\hat{\psi}) + R$$

$$\text{and } A(\hat{\psi}) = (I - \hat{\rho}W)^T (I - \hat{\rho}W).$$

This expression for the MSE of the EBLUP severely underestimates the true MSE as the variability due to the estimation of the parameters through the data has been ignored. We obtain a second order approximation to the $\text{MSE}[\hat{\theta}(\hat{\psi})]$ in case $\hat{\psi}$ is the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (REMLE) of ψ , with the assumption of large m and by neglecting all the terms of the order $o(m^{-1})$, under the following regularity conditions. The approximation has been worked out along the lines of Prasad and Rao (1990) and Datta and Lahiri (2000) which are heuristic in nature.

Regularity Conditions 1

- The elements of X are uniformly bounded such that $X^T \Sigma^{-1}(X)X = [O(m)]_{p \times p}$, where $\Sigma(\psi) = [\sigma_v^2 A^{-1}(\psi) + R]$;
- m is finite;
- $\Lambda(\psi)X = [O(1)]_{m \times p}$, $(\partial[\Lambda(\psi)X]) / (\partial\psi_d) = [O(1)]_{m \times p}$, $(\partial^2[\Lambda(\psi)]) / (\partial\psi_d \partial\psi_e) = [O(1)]_{m \times m}$ for $d, e = 1, 2$;
- $\hat{\psi}$ is the estimator of ψ which satisfies $\hat{\psi} - \psi = O_p(m^{-1/2})$, $\hat{\psi}(-y) = \hat{\psi}(y)$, $\hat{\psi}(y + xh) = \hat{\psi}(y) \forall h \in R^p$ and $\forall y$.

These regularity conditions are satisfied in this case. The special standardised form of the weight matrix W satisfies the condition (c) for $|\rho| < 1$ as it has only a finite number of nonzero elements and its row sum is equal to 1. It may be mentioned here that the matrix $\sigma_v^2 A^{-1} \Sigma^{-1}$ has finite number

of nonzero elements and the order of $W, (I - \rho W), W(I - \rho W), \Sigma, \Sigma^{-1}$ or any sum or product combination of these and their derivatives mentioned in condition (c) do not increase. The MLE and the REMLE, in addition satisfy the condition (d). A second order approximation to the MSE of the EBLUP has been shown in Theorem A.1 of the Appendix as

$$\text{MSE}[\hat{\theta}(\hat{\psi})] = E[(\hat{\theta}(\hat{\psi}) - \theta)(\hat{\theta}(\hat{\psi}) - \theta)^T] = g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (2.12)$$

Here the third term $g_3(\psi)$ comes from estimating the unknown parameter vector from the sample data and it is of the same order $O(m^{-1})$ as that of $g_2(\psi)$. Further $g_3(\psi)$ may be expressed as

$$g_3(\psi) = L^T(\psi) [I_{\psi}^{-1}(\psi) \otimes \Sigma(\psi)] L(\psi), \quad (2.13)$$

where

$$L(\psi) = \text{Col}[L_d(\psi)] = [L_p(\psi), L_{\sigma_v^2}(\psi)]^T,$$

$$L_d(\psi) = \frac{\partial \Lambda(\psi)}{\partial \psi_d}, d = 1, 2, \quad I_{\psi}(\psi) = E[-\frac{\partial^2 l}{\partial \psi \partial \psi^T}]$$

is the information matrix and \otimes represents Kronecker product. Further $g_3(\psi)$ may also be written as

$$g_3(\psi) = \sum_{d=1}^2 \sum_{e=1}^2 L_d(\psi) \Sigma(\psi) L_e^T(\psi) I_{de}^{-1}(\psi) \quad (2.14)$$

$$\text{where } I_{\psi}^{-1}(\psi) \equiv (I_{de}^{-1}(\psi)).$$

It is common practice to estimate the MSE of the EBLUP by replacing the unknown parameters including components of the variance by their respective estimators. This procedure can lead to severe underestimation of the true MSE (Prasad and Rao 1990, Singh, Stukel and Pfeffermann 1998). We obtain the estimator of the MSE of the EBLUP in Theorem A.2 of the Appendix for large m neglecting all terms of order $o(m^{-1})$. As a result we have the expressions

$$E[g_1(\hat{\psi}) + g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] = g_1(\psi) + o(m^{-1}), \quad (2.15)$$

$$E[g_2(\hat{\psi})] = g_2(\psi) + o(m^{-1})$$

$$\text{and } E[g_3(\hat{\psi})] = g_3(\psi) + o(m^{-1}), \quad (2.16)$$

and finally the estimator of the MSE of $\hat{\theta}(\hat{\psi})$ as

$$\text{mse}[\hat{\theta}(\hat{\psi})] = [g_1(\hat{\psi}) + g_2(\hat{\psi}) + 2g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] + o(m^{-1}), \quad (2.17)$$

$$\text{where } E[\text{mse}(\hat{\theta}(\hat{\psi}))] = \text{MSE}[\hat{\theta}(\hat{\psi})] + o(m^{-1}).$$

Obviously the additional terms, $g_3(\hat{\psi}), g_4(\hat{\psi})$ and $g_5(\hat{\psi})$ are the contributions, due to estimation of unknown parameter vector ψ by $\hat{\psi}$. The expressions for $g_4(\psi)$ and $g_5(\psi)$ up to order $o(m^{-1})$ are given by

$$g_4(\psi) = [b_{\psi}^T(\psi) \otimes I_m] \frac{\partial g_1(\psi)}{\partial \psi},$$

$$b_{\psi}(\psi) = \frac{1}{2} I_{\psi}^{-1}(\psi) \text{Col}_{1 \leq d \leq 2} \left[\text{Trace} \left[I_{\beta}^{-1}(\psi) \frac{\partial I_{\beta}(\psi)}{\partial \psi_d} \right] \right], \quad (2.18)$$

$$g_5(\psi) = \frac{1}{2} \text{Trace}_m \left[\frac{\partial^2 \Sigma(\psi)}{\partial \psi \partial \psi^T} [I_{\psi}^{-1}(\psi) \otimes (\Sigma^{-1}(\psi) R)] \right]. \quad (2.19)$$

Here $b_{\psi}(\psi)$ is the bias of $\hat{\psi}$ i.e., $E(\hat{\psi}) - \psi$ up to order $o(m^{-1})$ and $(\partial g_1(\psi))/(\partial \psi)$ is a partitioned matrix $[(\partial g_1(\psi))/(\partial \rho), (\partial g_1(\psi))/(\partial \sigma_v^2)]^T$ of order $(2m \times m)$ having 2 matrices of order $m \times m$ in a column. In the same way $(\partial^2 \Sigma(\psi))/(\partial \psi \partial \psi^T)$ is a partitioned matrix of order $(2m \times 2m)$ having 2 partitions, row and column wise with $(\partial^2 \Sigma(\psi))/(\partial \psi_d \partial \psi_e)$ being a general sub matrix of order $m \times m$ therein. $\text{Trace}(B) = \sum_{d=1}^2 B_{dd}$, where B is a square partitioned matrix with square sub matrices of similar order. In addition $g_4(\psi)$ and $g_5(\psi)$ may also be written as

$$g_4(\psi) = \frac{1}{2} \sum_{d=1}^2 \sum_{e=1}^2 I_{de}^{-1}(\psi) \text{Trace} \left[I_{\beta}^{-1}(\psi) \frac{\partial I_{\beta}(\psi)}{\partial \psi_d} \right] \frac{\partial g_1(\psi)}{\partial \psi_e}, \quad (2.20)$$

$$g_5(\psi) = \frac{1}{2} \sum_{d=1}^2 \sum_{e=1}^2 \left[R \Sigma^{-1}(\psi) \frac{\partial^2 \Sigma(\psi)}{\partial \psi_d \partial \psi_e} \Sigma^{-1}(\psi) R I_{de}^{-1}(\psi) \right]. \quad (2.21)$$

The expression (2.17) gives the matrix of the estimator of the MSE of EBLUP, $\hat{\theta}(\hat{\psi})$ and the MSE of the individual small area estimators may be obtained as the respective diagonal element. In case of simple model without the spatial autocorrelation, similar expressions can be obtained. In this case $g_5(\psi)$, however, becomes zero.

3. Spatial Temporal Model

In this section, State Space Models via Kalman filtering have been used to take the advantage of the time series data along with the common regression parameter and common autocorrelation parameter to strengthen the direct survey estimators at any point of time. This is especially advantageous in the case where the past survey estimates are more reliable. The models used in this category are the following

$$y_t = X_t \beta + Z v_t + \varepsilon_t, \quad \varepsilon_t \sim N_m(0, R_t), \quad Z = (I - \rho W)^{-1}, \quad (3.1)$$

$$v_t = \rho v_{t-1} + \eta_t, \quad \eta_t \sim N_m(0, \sigma_v^2 I) \quad t = 1, 2, \dots, T \quad \text{and} \quad \varepsilon_t \text{ and } \eta_t \text{ are independent of each other.} \quad (3.2)$$

Here the parameters have usual meaning as explained in the previous section. Weight matrix $W(m \times m)$ and design matrices $X_t(m \times p)$ are known, $Z(m \times m)$ is a matrix of coefficients of random effects and ρ is an unknown autocorrelation coefficient. R_t is a diagonal matrix of order m which may be expressed as $R_t = \text{diag}(\sigma_{1t}^2, \sigma_{2t}^2, \dots, \sigma_{mt}^2)$ where σ_{it}^2 's are known sampling variances corresponding to the i^{th} small area and t^{th} time point. β is unknown vector of fixed effects and $\psi = [\rho, \sigma_v^2, k]^T$ is a vector of three unknown parameters. These parameters are independent of time t . It may be noted that the random effects v_t have been allowed to change in accordance with (3.2) and k is temporal autoregressive parameter. For stationarity $|k| < 1$.

The estimators of fixed and random effects and the MSE of these estimators are obtained in stages, starting with assumption of mixed effects linear model approach at time $t = 1$, and by taking $v_1 \sim N_m(0, \sigma_v^2 I)$ (Sallás and Harville 1994). In the standard form we write the model as

$$y_t = U_t \alpha_t + \varepsilon_t, \quad \alpha_t = T \alpha_{t-1} + \zeta_t, \quad T = \text{diag}[I_p, k I_m], \quad (3.3)$$

$$\zeta_t \sim N_{p+m}(0, Q), \quad Q = \text{diag}[0_p, \sigma_v^2 I_m]$$

$$U_t = [X_t, Z], \quad \alpha_t = [\beta, v_t]^T. \quad (3.4)$$

Here I_m and 0_m are the unit and zero matrices of order m and by $\text{diag}[I_p, k I_m]$ we mean the matrix

$$\begin{bmatrix} I_{p \times p} & 0_{p \times m} \\ 0_{m \times p} & k I_{m \times m} \end{bmatrix}.$$

In case β is assumed fixed but dependent on time, there is no change in the model except that $T = \text{diag}[0_p, k I_m]$.

The initial estimates of the effects α_t and their variances (based on $t = 1$) are obtained as

$$\hat{\beta}_1 = (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} y_1, \quad \hat{v}_1 = \sigma_v^2 Z^T H_1^{-1} (y_1 - X_1 \hat{\beta}_1),$$

$$H_1 = R_1 \sigma_v^2 A^{-1}, \quad \Sigma_1 = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

$$\Sigma_{11}(p \times p) = (X_1^T H_1^{-1} X_1)^{-1},$$

$$\Sigma_{12}(p \times m) = \Sigma_{21}^T = -\sigma_v^2 (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} Z$$

$$\text{and } \Sigma_{22}(m \times m) = \sigma_v^2 I_m - \sigma_v^4 Z^T H_1^{-1} Z \\ + \sigma_v^4 Z^T H_1^{-1} X_1 (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} Z.$$

The recurring Kalman filtering equations for updation of the estimators at subsequent stages are

$$\begin{aligned}\Sigma_{t|t-1} &= T \Sigma_{t-1} T^T + Q, \hat{\alpha}_{t|t-1} = T \hat{\alpha}_{t-1}, H_t = R_t + U_t \Sigma_{t|t-1} U_t^T, \\ \hat{\alpha}_t &= \hat{\alpha}_{t|t-1} + \Sigma_{t|t-1} U_t^T H_t^{-1} (y_t - U_t \hat{\alpha}_{t|t-1}), \\ \Sigma_t &= \Sigma_{t|t-1} - \Sigma_{t|t-1} U_t^T H_t^{-1} U_t \Sigma_{t|t-1}\end{aligned}$$

where $\hat{\alpha}_{t|t-1}$ are the estimators of the effects α_t given the observations $[y_1, y_2, \dots, y_{t-1}]$ and the $\Sigma_{t|t-1}$ are the mean squared errors of $\hat{\alpha}_{t|t-1}$. H_t are the conditional variance covariance matrix of y_t given $[y_1, y_2, \dots, y_{t-1}]$. With the help of the above recurring filtering equations, the best linear unbiased predictor (BLUP) of $\theta_t = X_t \beta + Z_t v_t$, and the mean squared error (MSE) of the BLUP may be obtained as

$$\begin{aligned}\hat{\theta}_t(\psi) &= U_t(\psi) \hat{\alpha}_t(\psi) \\ &= y_t - R_t H_t^{-1}(\psi) [y_t - U_t(\psi) \hat{\alpha}_{t|t-1}(\psi)] \\ &= U_t(\psi) \hat{\alpha}_{t|t-1}(\psi) + \Lambda_t(\psi) e_t(\psi),\end{aligned}\quad (3.5)$$

$$\text{MSE}[\hat{\theta}_t(\psi)] = g_{12t}(\psi) = U_t(\psi) \Sigma_t(\psi) U_t^T(\psi), \quad (3.6)$$

$$\begin{aligned}\text{where } \Lambda_t(\psi) &= U_t(\psi) \Sigma_{t|t-1}(\psi) U_t^T(\psi) H_t^{-1}(\psi) \\ &= I_m - R_t H_t^{-1}(\psi) \\ \text{and } e_t(\psi) &= y_t - U_t(\psi) \hat{\alpha}_{t|t-1}(\psi).\end{aligned}$$

It may be noted that $g_{12t}(\psi)$ is the spatial counterpart of $g_1(\psi) + g_2(\psi)$. As usual in practice, the parameter vector ψ is unknown and its restricted maximum likelihood estimators (REMLE) can be obtained by maximizing the following log likelihood function, based on the sample data covering all time points

$$\begin{aligned}l = \text{const.} - \frac{1}{2} \log |X_1^T H_1^{-1} X_1| - \frac{1}{2} \sum_{t=1}^T \log |H_t| \\ - \frac{1}{2} (y_1 - X_1 \hat{\beta}_1)^T H_1^{-1} (y_1 - X_1 \hat{\beta}_1) \\ - \frac{1}{2} \sum_{t=2}^T (y_t - U_t \hat{\alpha}_{t|t-1})^T H_t^{-1} (y_t - U_t \hat{\alpha}_{t|t-1})\end{aligned}\quad (3.7)$$

with respect to the parameter ψ . With the help of the above, the estimator, $\hat{\psi}$ is obtained and the EBLUP of θ_t and the naive estimator of the MSE of the EBLUP are given by

$$\hat{\theta}_t(\hat{\psi}) = U_t(\hat{\psi}) \hat{\alpha}_t(\hat{\psi}) = U_t(\hat{\psi}) \hat{\alpha}_{t|t-1}(\hat{\psi}) + \Lambda_t(\hat{\psi}) e_t(\hat{\psi}), \quad (3.8)$$

$$\text{MSE}[\hat{\theta}_t(\hat{\psi})] = g_{12t}(\hat{\psi}) = U_t(\hat{\psi}) \Sigma_t(\hat{\psi}) U_t^T(\hat{\psi}). \quad (3.9)$$

As explained earlier in section 2, the MSE of the EBLUP underestimates the true MSE as it does not take care of the variability due to replacing parameters by their estimates. A second order approximation to the $\text{MSE}[\hat{\theta}_t(\hat{\psi})]$ for large m and neglecting all the terms of order $o(m^{-1})$, has been obtained in Theorem A.3 of the Appendix, under the

following regularity conditions satisfied by our model. These conditions are analogous to the regularity conditions 1.

Regularity Conditions 2

- The elements of $X_t, t=1, 2, \dots, T$ are uniformly bounded such that $X_t^T \Sigma_t^{-1}(\psi) X_t = [O(m)]_{p \times p}$, where $\Sigma_t(\psi) = [\sigma_v^2 A^{-1}(\psi) + R_t]$;
- m and T are finite;
- $\Lambda_t(\psi) U_t(\psi) = [O(1)]_{m \times p}$, $(\partial \Lambda_t(\psi) U_t(\psi)) / (\partial \psi_d) = [O(1)]_{m \times p}$, $([\partial^2 \Lambda_t(\psi)] / (\partial \psi_d \partial \psi_e)) = [O(1)]_{n \times m}$, $t=1, 2, \dots, T$ and $d, e=1, 2, 3$;
- $\hat{\psi}$ is the estimator of ψ which satisfies $\hat{\psi} - \psi = O_p(m^{-1/2})$, $\hat{\psi}(-y) = \hat{\psi}(y)$, $\hat{\psi}(y + xh) = \hat{\psi}(y) \forall h \in R^p$ and $\forall y$.

The second order approximation to the MSE of the EBLUP is

$$\begin{aligned}\text{MSE}[\hat{\theta}_t(\hat{\psi})] &= E[(\hat{\theta}_t(\hat{\psi}) - \theta_t)(\hat{\theta}_t(\hat{\psi}) - \theta_t)^T] \\ &= g_{12t}(\psi) + g_{3t}(\psi) + o(m^{-1}).\end{aligned}\quad (3.10)$$

Here $g_{3t}(\psi)$ is the bias due to the estimation of the parameters from the sample data and is of the order $O(m^{-1})$ and it is given by

$$g_{3t}(\psi) = L_t^T(\psi) I_\psi^{-1}(\psi) K_\psi(\psi) H_t I_\psi^{-1}(\psi) L_t(\psi) \quad (3.11)$$

where $K_\psi(\psi) \equiv (K_{de}(\psi))$

$$\text{and } K_{de}(\psi) = \frac{1}{2} \sum_{i=1}^T \text{Trace} \left[H_i^{-1} \frac{\partial H_i}{\partial \psi_d} H_i^{-1} \frac{\partial H_i}{\partial \psi_e} \right]. \quad (3.12)$$

Further

$$L_t(\psi) = \text{Col} [L_{td}(\psi)] \text{ and } L_{td}(\psi) = (\partial \Lambda_t(\psi)) / (\partial \psi_d)$$

for $d=1, 2, 3$.

In a proper form, we may write $g_{3t}(\psi)$ as

$$\begin{aligned}g_{3t}(\psi) &= \left[\sum_{f=1}^3 \sum_{g=1}^3 I_{df}^{-1}(\psi) \right. \\ &\quad \times \sum_{i=1}^T \text{Trace} \left(H_i^{-1} \frac{\partial H_i}{\partial \psi_f} H_i^{-1} \frac{\partial H_i}{\partial \psi_g} \right) L_{te}^T(\psi) \\ &\quad \times H_t I_{ge}^{-1}(\psi) \end{aligned}$$

The expression for the information matrix involved here, may be given as

$$\begin{aligned}
I_{de}(\psi) &= E \left[-\frac{\partial^2 l}{\partial \psi_d \partial \psi_e} \right] \\
&= \frac{1}{2} \sum_{t=1}^T \text{Trace} \left[H_t^{-1} \frac{\partial H_t^{-1}}{\partial \psi_d} H_t^{-1} \frac{\partial H_t^{-1}}{\partial \psi_e} \right] + \sum_{t=1}^T \left[\frac{\partial e_t^T}{\partial \psi_d} H_t^{-1} \frac{\partial e_t}{\partial \psi_e} \right] \\
&\quad - \frac{1}{2} \text{Trace} \left[\left(X_1^T H_1^{-1} X_1 \right)^{-1} X_1^T H_1^{-1} \right. \\
&\quad \times \left(\frac{\partial^2 H_1}{\partial \psi_d \partial \psi_e} - 2 \frac{\partial H_1}{\partial \psi_d} H_1^{-1} \frac{\partial H_1}{\partial \psi_e} \right) H_1^{-1} X_1 \left. \right] \\
&\quad - \frac{1}{2} \text{Trace} \left[\left(X_1^T H_1^{-1} X_1 \right)^{-1} X_1^T H_1^{-1} \frac{\partial H_1}{\partial \psi_d} H_1^{-1} X_1 \right. \\
&\quad \times \left(X_1^T H_1^{-1} X_1 \right)^{-1} X_1^T H_1^{-1} \frac{\partial H_1}{\partial \psi_e} H_1^{-1} X_1 \left. \right].
\end{aligned}$$

Estimator of the MSE of the EBLUP has also been obtained with the assumption of large m and neglecting all terms of order $o(m^{-1})$ in Theorem A.4 of the Appendix as

$$\begin{aligned}
\text{mse}(\hat{\theta}_t(\hat{\psi})) &= [g_{12t}(\hat{\psi}) + g_{3t}(\hat{\psi}) + g_{31t}(\hat{\psi}) \\
&\quad - g_{4t}(\hat{\psi}) - g_{5t}(\hat{\psi})] + o(m^{-1}), \quad (3.13)
\end{aligned}$$

where $g_{31t}(\psi)$, $g_{4t}(\psi)$ and $g_{5t}(\psi)$ are given as

$$g_{31t}(\psi) = L_t^T(\psi) [I_\psi^{-1}(\hat{\psi}) \otimes H_t(\psi)] L_t(\psi), \quad (3.14)$$

$$\begin{aligned}
g_{4t}(\psi) &= [b_\psi^T(\psi) \otimes I_m] \frac{\partial g_{12t}(\psi)}{\partial \psi}, \\
b_\psi &= \frac{1}{2} I_\psi^{-1}(\psi) \text{Col}_{1 \leq d \leq 3} \left[\text{Trace} \left[I_\beta^{-1}(\psi) \frac{\partial I_\beta(\psi)}{\partial \psi_d} \right] \right], \quad (3.15)
\end{aligned}$$

$$\begin{aligned}
g_{5t}(\psi) &= \\
&\quad \frac{1}{2} \text{Trace}_m \left[\frac{\partial^2 H_t}{\partial \psi \partial \psi^T} [I_\psi^{-1}(\psi) \otimes (H_t^{-1} R_t)] \right]. \quad (3.16)
\end{aligned}$$

4. Analysis of the NSSO Data

National Sample Survey Organisation (NSSO) of the Ministry of Statistics and Programme Implementation (Government of India) conducts quinquennial large sample surveys (QS) on household consumption expenditure and employment, almost every five years in India. The surveys cover more than hundred thousand households spread over a number of villages and urban blocks. In order to fill the gaps in data between the successive QSSs, the NSSO conducts annual consumer expenditure survey (CES) in almost every round (equivalent to six months or one year duration). The annual series covers only 10–30 thousand households depending on the number of villages and urban blocks surveyed all over the country. Each round of NSS normally

has more than one subject of enquiry. The annual series has a different principal subject of enquiry. However schedule 1.0 of the annual surveys is designed to collect data on household consumption expenditure among other characteristics on employment.

The NSSO adopts two stage stratified sampling design, the first stage units being census villages in the rural sector selected through circular systematic sampling with probability proportional to size (PPS) and the ultimate-stage units being the households selected circular systematically with independent random starts. India has been divided into States and the Districts are the second level administrative units in the States. There is not much difference between the annual and quinquennial surveys excepting that normally in annual series, a small sample of four households per first stage units are surveyed while in the case of quinquennial survey, ten to twelve households per first stage units are surveyed. Besides this, in NSSO surveys, we have two samples viz, the first one as central sample surveyed by the investigators of the NSSO, and the second one as state sample surveyed by the State authorities. Regarding the estimation procedure, the first stage units are selected in the form of two independent sub-samples. The estimate of the population mean and its variance based on the two sub-samples are separately obtained. The pooled mean $y_i = (\hat{y}_{1i} + \hat{y}_{2i})/2$ and $R_i = (\hat{y}_{1i} - \hat{y}_{2i})^2/4$ for $i = 1, 2, \dots, m$, where \hat{y}_{1i} , \hat{y}_{2i} are the sub-sample means, estimate respectively the population mean and its variance for a particular district (small area). In case of round 55, first stage units are selected in the form of eight independent sub-samples and the estimate of the population mean and its variance are based on these sub-samples. In view of the problems related to the estimates of R_i 's with 1 d.f., the R_i for each small area were analysed and compared over time. In case of any abnormal R_i , it was smoothed out by taking the average of R_i 's over neighboring time points and in some cases, over neighboring small areas also. The survey estimates y_i 's are the direct estimates, and the smoothed R_i 's are the diagonal elements of the sampling variance covariance matrix R , in our model equations (2.1), (2.4) and (3.1), referred in this paper.

In this paper, we have used data from central sample only. The estimates of monthly per capita consumption expenditure (MPCE) and of the standard errors (SE) of the estimators have been obtained under various mixed effects models for the rural 63 districts (small areas) of a large state in India, namely, Uttar Pradesh. We have used data from the six rounds of the NSSO viz round 50 (July 1993–June 1994), round 51 (July 1994–June 1995), round 52 (July 1995–June 1996), round 53 (January–December 1997), round 54 (January–June 1998) and round 55 (July 1999–June 2000). Out of these rounds 50 and 55 are based on

quinquennial surveys. The selected exogenous variables used in the models are i) number of households, ii) gross area sown and iii) per capita net area sown in the districts. The agricultural data are available on annual basis while the estimates of the households and the population were obtained through the interpolation techniques based on the 1971, 1981 and 1991 decennial census data. These exogenous variables have been selected from a host of variables ranging from 1991 census to annual agricultural data through the covariate analysis. Different weight matrices such as length of common boundary between a pair of districts, distance between centres of two districts and the binary weights were considered. Binary weights give larger estimate of spatial autocorrelation coefficient, therefore they (standardised by making row sum of the weight matrix as one) have been used for further analysis in this paper. In the whole exercise, maximization of log likelihood function and the estimation of the parameters have been carried out by using the Nelder and Mead simplex method on the software MATLAB.

Various mixed effects models, used for finding out improved estimates of MPCE are given in Table 1. The parameters in the models have usual meaning as shown in sections 2 and 3. Further, in case of each model, sampling variance R or R_t (in case of temporal model) are assumed to be known.

Table 1
Mixed Effects Models

Model-1	Direct Estimates	
Model-2	Regression Model	$y = X\beta + v + \varepsilon$
Model-3	Spatial Model	$y = X\beta + Zv + \varepsilon$
Model-3A	Spatial Model (intercept)	$y = \mu + Zv + \varepsilon$
Model-4	Regression Temporal	$y_t = X_t\beta + v_t + \varepsilon_t, v_t = kv_{t-1} + \eta_t$
Model-5	Spatial Temporal	$y_t = X_t\beta + Zv_t + \varepsilon_t, v_t = kv_{t-1} + \eta_t$

Table 2 presents the round wise estimates of the parameters for the simple mixed effects regression and spatial models. The value of the multiple correlation coefficients R^2 between MPCE estimates and the auxiliary variables, in case of each round has also been shown here. The figures in bracket show the Standard Errors (SE) of the parameter estimates. Note that $\lambda(=\lambda_1, \lambda_2)$ is the likelihood ratio test (LRT) statistics defined as $-2 \log L \sim \chi_k^2$, where L is the ratio of nested likelihoods at the hypothesised parameter values for two competing models under different hypotheses and k is the difference between the number of parameters under two models. Here λ_1 compares regression model and spatial model, under $H_0: \rho = 0$ against $H_1: \rho \neq 0$ and is distributed as χ_1^2 under H_0 , and λ_2 compares spatial model and spatial (intercept) model, under $H_0: \beta = 0$ against $H_1: \beta \neq 0$ [β does not include intercept term β_0] and is distributed as χ_3^2 under H_0 .

On comparison of the simple regression model (Model 2) and spatial model (Model 3) through LRT, we find that under $H_0(\rho = 0)$, the spatial autocorrelation ρ for Model 3 has been found highly significant for the two rounds 52 and 55, obviously for these rounds, use of spatial model results in much improvement in the estimates of MPCE. On the other hand, in case of rounds 50 and 53, and for these only, the regression coefficients β have been found nearly significant for the Model 3 in comparison to Model 3A which shows that the spatial model with intercept term may improve the estimates for these rounds without any help of the exogenous variables.

Table 3 presents the parameter estimates and their SE in case of regression temporal model and spatial temporal model.

For Model 4, unconstrained iterative maximisation process converged the value of k greater than 1, which is inadmissible under the assumption of stationarity. For this

Table 2
Estimates of Parameters for Small Area Estimates of MPCE Under Regression and Spatial Models

Round	R^2	Model 2	Model 3	LRT	Model 3A	LRT
		σ_v^2	ρ	σ_v^2	ρ	σ_v^2
Rd. 50	0.27	1,724.48 (356.19)	0.30 (0.18)	1,635.70 (346.45)	1.80	0.59 (0.13)
Rd. 51	0.27	3,424.21 (820.89)	0.48 (0.19)	3,156.90 (815.24)	0.66	0.67 (0.13)
Rd. 52	0.17	2,150.54 (540.23)	0.87 (0.07)	714.96 (237.15)	13.46	0.86 (0.07)
Rd. 53	0.13	6,312.99 (1,397.92)	-0.39 (0.27)	5,822.99 (1,374.70)	1.56	0.09 (0.23)
Rd. 54	0.22	3,437.67 (806.87)	0.61 (0.14)	2,793.24 (742.35)	1.30	0.66 (0.13)
Rd. 55	0.31	2,989.73 (712.28)	0.87 (0.06)	1,060.21 (362.40)	20.30	0.86 (0.07)

λ_1 and λ_2 compare models 2,3 and models 3,3A respectively. $\chi_{1,05}^2 = 3.841$ for λ_1 and $\chi_{3,05}^2 = 7.815$ for λ_2 .

case, estimates were obtained by taking $k=1$ and Model 4 was accordingly modified. Table 3 reports the results for $k=1$ in case of regression temporal model. The spatial temporal model shows higher value of common autocorrelation coefficient and far lower value of the estimate of σ_v^2 . A summary of the round wise average estimates of MPCE (based on all the 63 districts), their estimated standard errors (SE) and the coefficient of variation (CV) under each model has been presented in Table 4.

The results of Table 4 have been summarized below.

The Direct survey estimates are less precise and all the models involving mixed effects improve it. The estimates for the rounds 50 and 55 (based on large samples) are more precise than the estimates based on other rounds. Spatial model, depending on the value of ρ improves the estimates considerably. In case of rounds 52 and 53, where the autocorrelation have been found significant, the reduction in the average SE of the estimates in comparison to the model without spatial autocorrelation, is considerable. Model 3A with spatial effect and without auxiliary variables is equally

good. The spatial temporal model further improves the estimates taking into advantage of the state space considerations. It may be noted that for the round 52 (very high spatial autocorrelation), the estimates based on temporal models are worse than the estimates based on models without temporal considerations. Perhaps due to fixed regression and autocorrelation parameters, the estimates tend towards the average of the five rounds.

In order to judge the performances of the estimators under various models vis-a-vis under the most general model (spatial temporal model), data have been simulated under the spatial temporal model and true MSEs of the replicated estimates under each of the assumed models have been obtained. For this, we have conducted the simulation by taking the estimated parameters from the spatial temporal model, given in Table 2 and obtained the true replicated small area mean $\theta(b)$ for b^{th} replication ($b=1, 2, \dots, B$) along with simulated observations $y(b)$ for a large number of replications. On this simulated dataset, for each replication, different models including spatial temporal model

Table 3
Estimates of Parameters for Small Area Estimation of MPCE Under Regression Temporal and Spatial Temporal Models

Models	ρ		σ_v^2		k	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Model 4	—	—	4,715.64	431.00	—	—
Model 5	0.79	0.04	2,163.50	245.50	0.53	0.07

Table 4
Average EBLUP for MPCE (Rs.), their Estimated SE and CV Under Regression, Spatial, Regression Temporal and Spatial Temporal Models

Models	NSSO Rounds					
	50	51	52	53	54	55
Average Small Area Estimates						
Model 1	276.10	321.26	373.07	408.52	411.25	482.00
Model 2	272.87	312.53	354.45	397.52	400.87	471.99
Model 3	272.98	313.14	351.51	398.21	400.78	471.09
Model 3A	273.56	314.19	352.01	396.40	399.91	471.91
Model 4	274.13	305.62	345.54	383.53	399.56	463.32
Model 5	273.75	312.21	351.79	391.61	399.50	473.57
Average Standard Errors (SE)						
Model 1	25.09	66.06	64.18	74.19	53.87	45.45
Model 2	17.10	33.65	29.09	39.85	32.68	30.59
Model 3	16.88	32.84	21.51	39.98	30.87	24.84
Model 3A	16.56	31.29	20.79	40.03	30.23	24.37
Model 4	19.51	34.91	35.19	37.79	35.14	33.15
Model 5	17.18	28.99	28.33	30.02	28.76	28.10
Average Coefficient of Variation (CV) (%)						
Model 1	9.09	20.56	17.20	18.16	13.10	9.43
Model 2	6.27	10.79	8.21	10.01	8.15	6.48
Model 3	6.18	10.49	6.12	10.04	7.70	5.27
Model 3A	6.05	9.96	5.91	10.10	7.56	5.17
Model 4	7.12	11.42	10.18	9.85	8.79	7.15
Model 5	6.28	9.29	8.05	7.67	7.20	5.93

Table 5
Percentage Relative Efficiency [RMSE] of the Temporal Models in Comparison to other Models for MPCE

	NSSO Rounds					
	50	51	52	53	54	55
Spatial Temporal Model [Model 5]						
Model 2	123.63	170.54	193.68	203.55	204.72	169.76
Model 3	100.24	133.82	149.70	165.46	165.85	154.23
Model 4	125.81	141.50	141.93	137.55	139.11	129.88
Regression Temporal Model [Model 4]						
Model 2	100.71	134.50	156.35	165.30	163.13	152.56

have been applied and the small area mean estimators under each of them are obtained. While fitting the regression and spatial temporal models on the simulated datasets, the iterative maximisation process have the constrained value of $k \leq 1$. Here we have taken $B = 5,000$ replications. The true MSEs of the estimators for i^{th} small area under a particular model ($k = 2-4$) may be defined as

$$\text{MSE}(\theta_i^k) = \frac{1}{B} \sum_{k=1}^B [\hat{\theta}_i^k(b) - \theta_i(b)]^2, \quad i = 1, 2, \dots, m.$$

The relative efficiency of the estimators under spatial temporal model (Model 5) against the estimators under models 2–4 have been judged by the ratio of their mean squared errors (RMSE) as

$$\text{RMSE}(k, \text{Temp}) = 100 \frac{\sum_{i=1}^m \text{MSE}(\hat{\theta}_i^k)}{\sum_{i=1}^m \text{MSE}(\hat{\theta}_i^{\text{Temp}})}$$

where ‘Temp’ denotes the spatial temporal model and k denotes models 2, 3 and 4. Likewise the relative efficiency of the regression temporal model (Model 4) against the simple regression model (Model 2) has been found by simulating data with the estimated parameters given in Table 3, under the regression temporal model. The results have been shown in Table 5.

The results confirm the superiority of the spatial temporal model in comparison to other models for these parameters. The regression temporal model has also been found better than the simple regression model.

5. Conclusions

The Direct survey estimates based on the small sample can be considerably improved by using the area specific small area models. The spatial autocorrelation amongst the neighboring areas may be exploited for improving the direct survey estimates. However, the model must be applied after studying the significant correlation amongst the small areas by virtue of their neighborhood effects. In case of poor relation between the dependent and exogenous variables, the simple spatial model with intercept only, may equally

improve the estimates. This model uses only the spatial autocorrelation to strengthen the small area estimates and do not require the use of exogenous variables. The spatial models, by using the appropriate weight matrix W , or a combination of W matrices, can considerably improve the estimates. Weight matrix should be based on logical considerations and it may be used effectively for the cases, where due to some reasons, reliable exogenous variables are not available. This aspect can be further exploited to find out the small area estimates for the areas which have been recently created/demarcated.

One has to be careful about the increase in the MSE due to the variability caused by replacing the parameters by their estimates. This gets reflected through the second order approximation to the MSE dealt in the paper. That is why many times the simple spatial model (with intercept) performs better than the spatial model involving more parameters. Use of time series data with fixed regression parameters across the time, further improves the small area estimates especially for the time points where the direct survey estimates have larger MSE. Spatial temporal models have advantage over temporal models without spatial consideration due to the inclusion of fixed spatial autocorrelation across the small areas. However, for some time points for which ρ may be very different than the rest, this may not hold due to estimates tending towards the average of five rounds. Here the temporal consideration can be started from a suitable initial time point. Finally the exogenous variables X and the weight matrix W supplement each other through the regression parameter β and the autocorrelation parameter ρ and a judicious use of them may result in considerable improvement in the small area estimates.

Acknowledgements

The unit level data for the research have been made available by the National Sample Survey Organisation (NSSO), Ministry of Statistics and Programme Implementation under a research collaboration between IIT Kanpur and the NSSO. The weight matrix containing the length of

the boundary between different small areas (districts) have been provided by the National Informatics Centre (NIC) of the Ministry of Information Technology, Government of India. We would like to thank the referees for their helpful comments which has considerably improved the paper.

Appendix

Theorem A.1: Under Regularity Conditions 1

$$\text{MSE}[\hat{\theta}(\psi)] = g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (5.1)$$

For proof of the Theorem, we use the following well known results (Srivastawa and Tiwari 1976). Let $U \sim N(0, \Sigma)$ then for the symmetric matrices A, B and C

$$\begin{aligned} E[U(U^T A U)U^T] &= \text{Trace}(A\Sigma)\Sigma + 2\Sigma A\Sigma \\ E[U(U^T A U)(U^T B U)U^T] &= \text{Trace}(A\Sigma)\text{Trace}(B\Sigma)\Sigma \\ &+ 2[\text{Trace}(A\Sigma)\Sigma B\Sigma + \text{Trace}(B\Sigma)\Sigma A\Sigma + \text{Trace}(A\Sigma B\Sigma)\Sigma] \\ &+ 4[\Sigma A\Sigma B\Sigma + \Sigma B\Sigma A\Sigma]. \end{aligned}$$

Proof of Theorem A.1

Kackar and Harville (1984) showed that $\text{MSE}[\hat{\theta}(\psi)] = \text{MSE}[\hat{\theta}(\psi)] + E[(\hat{\theta}(\psi) - \hat{\theta}(\psi))(\hat{\theta}(\psi) - \hat{\theta}(\psi))^T]$. It is straight forward to show that $\text{MSE}[\hat{\theta}(\psi)] = g_1(\psi) + g_2(\psi)$. We need to prove that $g_3(\psi) = E[(\hat{\theta}(\psi) - \hat{\theta}(\psi))(\hat{\theta}(\psi) - \hat{\theta}(\psi))^T] + o(m^{-1})$. Taylor Series expansion of $\hat{\theta}(\psi)$ around ψ and using $(\hat{\psi} - \psi) = O_p(m^{-1/2})$ and $(\partial^2 \hat{\theta}(\psi))/(\partial \psi_d \partial \psi_e)|_{\psi=\hat{\psi}} = O_p(1)$ when $\|\hat{\psi}^* - \hat{\psi}\| \leq \|\hat{\psi} - \psi\|$ we get

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = [(\hat{\psi} - \psi) \otimes I_m]^T \nabla \hat{\theta}(\psi) + O_p(m^{-1}). \quad (5.2)$$

Here $\nabla \hat{\theta}(\psi) = (\partial \hat{\theta}(\psi))/(\partial \psi) = [(\partial \hat{\theta}(\psi))/(\partial \rho), (\partial \hat{\theta}(\psi))/(\partial \sigma_v^2)]^T$. Using

$$\begin{aligned} \frac{\partial \hat{\theta}(\psi)}{\partial \psi_d} &= \sum_{\alpha=1}^p \frac{\partial \hat{\theta}^*(\beta, \psi)}{\partial \beta_\alpha} \Big|_{\beta=\beta(\psi)} \frac{\partial \beta(\psi)}{\partial \psi_d} + \frac{\partial \hat{\theta}^*(\beta, \psi)}{\partial \psi_d} \Big|_{\beta=\beta(\psi)} \\ d &= 1, 2 \end{aligned}$$

where $\hat{\theta}^*(\beta, \psi) = X\beta(\psi) + \Lambda(\psi)[y - X\beta(\psi)]$, and the fact that $(\partial \beta_\alpha(\psi))/(\partial \psi_d) = O_p(m^{-1/2})$ (Cox and Reid (1987)), we get from the above

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = [(\hat{\psi} - \psi)^T \otimes I_m] \nabla \hat{\theta}^*(\psi) + O_p(m^{-1}) \quad (5.3)$$

$$\begin{aligned} \text{where } \nabla \hat{\theta}^*(\psi) &= \left[\frac{\partial \hat{\theta}^*(\beta, \psi)}{\partial \rho}, \frac{\partial \hat{\theta}^*(\beta, \psi)}{\partial \sigma_v^2} \right]^T \Big|_{\beta=\beta(\psi)} \\ &= L(\psi)[y - X\hat{\beta}(\psi)]. \end{aligned}$$

Using the Regularity Conditions 1 and the fact that $\hat{\beta}(\psi) - \beta = O_p(m^{-1/2})$ we have

$$\begin{aligned} [\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] &= [(\hat{\psi} - \psi) \otimes I_m]^T L(\psi)[y - X\hat{\beta}(\psi)] + O_p(m^{-1}) \\ &= \sum_{d=1}^2 (\hat{\psi}_d - \psi_d) L_d(\psi)[y - X\hat{\beta}(\psi)] + O_p(m^{-1}). \end{aligned}$$

Further using the Taylor Series expansion of the Likelihood $S(\hat{\eta}) = 0$ around ψ where

$$S(\eta) = [S_\beta^T(\eta), S_\psi^T(\eta)]^T, S_\beta^T(\eta) = \text{Col}_{1 \leq \alpha \leq p} \left[\frac{\partial \ell}{\partial \beta_\alpha} \right]$$

and the orthogonality of β and ψ , it follow that

$$(\hat{\psi} - \psi) = I_\psi^{-1}(\psi) S_\psi(\eta) + O_p(m^{-1}).$$

Writing

$$\begin{aligned} S_\psi(\psi) &= \text{Col}_{1 \leq d \leq 2} [S_d(\psi)] = [S_\rho(\psi), S_{\sigma_v^2}(\psi)]^T, \\ S_d(\psi) &= \frac{\partial \ell}{\partial \psi_d} = -\frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \right] + \frac{1}{2} [u^T B_d(\psi) u], \end{aligned}$$

$$B_d(\psi) = \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1}, u = y - X\beta(\psi) \text{ and}$$

$$I_{de}(\psi) = \frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \right]$$

we get

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = L^T(\psi) [I_\psi^{-1}(\psi) \otimes I_m] [S_\psi(\psi) \otimes u]$$

and thus the expression

$$\begin{aligned} &[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)][\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)]^T \text{ upto order } o(m^{-1}) \\ &= L^T(\psi) [I_\psi^{-1}(\psi) \otimes I_m] \text{Col}_{1 \leq d \leq 2} [u S_d(\psi)] \text{Concat}_{1 \leq e \leq 2} [S_e(\psi) u^T] \\ &\quad [I_\psi^{-1}(\psi) \otimes I_m] L(\psi) \\ &= L^T(\psi) [I_\psi^{-1}(\psi) \otimes I_m] \text{Col}_{1 \leq d \leq 2} \text{Concat}_{1 \leq e \leq 2} [u S_d(\psi) S_e(\psi) u^T] \\ &\quad [I_\psi^{-1}(\psi) \otimes I_m] L(\psi). \quad (5.4) \end{aligned}$$

Now we can write the likelihood and its derivative as

$$\begin{aligned} \ell &= \log L = \text{const.} - \frac{1}{2} \log |\Sigma| - \frac{1}{2} u^T \Sigma^{-1} u \\ \frac{\partial \ell}{\partial \psi_d} &= -\frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \right] + \frac{1}{2} u^T B_d(\psi) u, \end{aligned}$$

$$B_d(\psi) = \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1}$$

$$E \left[-\frac{\partial^2 \ell}{\partial \psi_d \partial \psi_e} \right] = \frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \right] = I_{de}(\psi)$$

where information matrix $I_\psi(\psi) \equiv I_{de}(\psi)$.

The expectation of a typical element of the inner most terms in the expression (5.4) becomes

$$E[uS_d(\psi)S_e(\psi)u^T] = E \begin{bmatrix} u[u^T B_d(\psi)u][u^T B_e(\psi)u]u^T \\ -u \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \right] [u^T B_e(\psi)u]u^T \\ -u[u^T B_d(\psi)u] \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \right] u^T \\ +u \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \right] \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \right] u^T \end{bmatrix}$$

and by applying the results of Srivastawa and Tiwari (1976), it becomes

$$E[uS_d(\psi)S_e(\psi)u^T] = \frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \right] \Sigma + 2 \left[\frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \right].$$

Substituting these in the expression (5.4) and also the second expression being of order $O(m^{-1})$, we can get the following upto order $o(m^{-1})$

$$\begin{aligned} & [\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)][\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)]^T \\ &= L^T(\psi)[I_{\psi}^{-1}(\psi) \otimes I_m] \text{Col Concat} [I_{de}(\psi)\Sigma] \\ & \quad [I_{\psi}^{-1}(\psi) \otimes I_m] L(\psi) \\ &= L^T(\psi)[I_{\psi}^{-1}(\psi) \otimes I_m][I_{\psi}(\psi) \otimes \Sigma][I_{\psi}^{-1}(\psi) \otimes I_m] L(\psi) \\ &= L^T(\psi)[I_{\psi}^{-1}(\psi) \otimes \Sigma] L(\psi). \end{aligned}$$

Theorem A.2: Under Regularity Conditions 1

$$E[g_1(\hat{\psi}) + g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] = g_1(\psi) + o(m^{-1}), \quad (5.5)$$

$$\begin{aligned} E[g_2(\hat{\psi})] &= g_2(\psi) + o(m^{-1}), \\ E[g_3(\hat{\psi})] &= g_3(\psi) + o(m^{-1}) \end{aligned} \quad (5.6)$$

$$\text{and } E[g_5(\hat{\psi})] = g_5(\psi) + o(m^{-1}). \quad (5.7)$$

Proof of Theorem A.2

Taylor Series expansion of $g_1(\hat{\psi})$ around ψ and using $\hat{\psi} - \psi = O_p(m^{-1/2})$ when $\|\hat{\psi}^* - \psi\| \leq \|\hat{\psi} - \psi\|$, we get

$$\begin{aligned} g_1(\hat{\psi}) &= g_1(\psi) + [(\hat{\psi}) - (\psi)]^T \otimes I_m] \nabla g_1(\psi) \\ & \quad + \frac{1}{2} [(\hat{\psi} - \psi)^T \otimes I_m] \nabla^2 g_1(\psi) [(\hat{\psi} - \psi) \otimes I_m] \\ & \quad + o_p(m^{-1}) \\ \nabla g_1(\psi) &= \left[\frac{\partial g_1(\psi)}{\partial \rho} \quad \frac{\partial g_1(\psi)}{\partial \sigma_v^2} \right]^T, \\ \nabla^2 g_1(\psi) &= \text{Col}_{1 \leq d \leq 2} \left[\text{Concat}_{1 \leq e \leq 2} \frac{\partial^2 g_1(\psi)}{\partial \psi_d \partial \psi_e} \right] \\ \frac{\partial g_1(\psi)}{\partial \psi_d} &= R \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} R \\ \frac{\partial^2 g_1(\psi)}{\partial \psi_d \partial \psi_e} &= -2R \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \Sigma^{-1} R \\ & \quad + R \Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \psi_d \partial \psi_e} \Sigma^{-1} R. \end{aligned}$$

Using the fact that $\Sigma(\psi)$ and its derivatives are symmetric, we have the second term of the expression as

$$\begin{aligned} & [(\hat{\psi} - \psi)^T \otimes I_m] \nabla^2 g_1(\psi) [(\hat{\psi} - \psi) \otimes I_m] \\ &= -L^T(\psi)[I_{\psi}^{-1}(\psi) \otimes \Sigma] L(\psi) \\ & \quad + \frac{1}{2} \text{Trace}_m \left[[I_2 \otimes (R \Sigma^{-1})] \frac{\partial^2 \Sigma}{\partial \psi_d \partial \psi_e} [I_{\psi}^{-1}(\psi) \otimes (\Sigma^{-1} R)] \right] \\ &= -g_3(\psi) + g_5(\psi) \end{aligned}$$

where $I_{\psi}^{-1}(\psi) = \text{Var}(\psi)$ is information matrix, the asymptotic variance of ψ . The first term in the expression $[(\hat{\psi} - \psi)^T \otimes I_m] \nabla g_1(\psi)$ reduces to $g_4(\psi)$ because of $E(\hat{\psi} - \psi) = b_{\hat{\psi}}(\psi)$ up to order $o(m^{-1})$ (Peers and Iqbal 1985).

The second part of the Theorem follows from the Taylor series expansion of $g_2(\hat{\psi})$, $g_3(\hat{\psi})$ and $g_5(\hat{\psi})$, each around ψ and using $\hat{\psi} - \psi = O_p(m^{-1/2})$ and $(\partial^2 g_2(\psi)) / (\partial \psi_d \partial \psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$, $(\partial^2 g_3(\psi)) / (\partial \psi_d \partial \psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$ and $(\partial^2 g_5(\psi)) / (\partial \psi_d \partial \psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$, respectively where $\|\hat{\psi}^* - \psi\| \leq \|\hat{\psi} - \psi\|$.

Theorem A.3: Under Regularity Conditions 2

$$\text{MSE}(\hat{\theta}_t(\hat{\psi})) = g_{12t}(\psi) + g_{3t}(\psi) + o(m^{-1}). \quad (5.8)$$

Proof of Theorem A.3

The proof is basically on the line of Theorem A.1 and with the use of the results of (Srivastawa and Tiwari (1976)) mentioned therein.

$$\begin{aligned} & \text{MSE}([\hat{\theta}_t(\hat{\psi})]) \\ &= \text{MSE}([\hat{\theta}_t(\psi)] + E[(\theta_t(\psi) - \theta_t)(\theta_t(\psi) - \theta_t)]^T] \\ &= g_{12t}(\psi) + E[(\theta_t(\psi) - \theta_t)(\theta_t(\psi) - \theta_t)]^T. \end{aligned} \quad (5.9)$$

Taylor series expansion of $\theta_t(\psi)$ around ψ and using $(\hat{\psi} - \psi) = O_p(m^{-1/2})$ and $(\partial^2 \theta(\psi)) / (\partial \psi_d \partial \psi_e) \big|_{\psi=\hat{\psi}^*} = O_p(1)$ when $\|\hat{\psi}^* - \psi\| \leq \|\hat{\psi} - \psi\|$ we have

$$\begin{aligned} & [\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi)] \\ &= [(\hat{\psi} - \psi) \otimes I_m]^T \nabla \hat{\theta}_t(\psi) + O_p(m^{-1}) \\ &= \sum_{d=1}^3 [(\hat{\psi}_d - \psi_d) L_{id}(\psi) e_t(\psi)] + O_p(m^{-1}). \end{aligned} \quad (5.10)$$

Further using the Taylor series expansion of the Likelihood equation $S(\hat{\eta}) = 0$ and the orthogonality of β and ψ , it follows

$$(\hat{\psi} - \psi) = I_\psi^{-1}(\psi) S(\psi) + O_p(m^{-1}). \quad (5.11)$$

Substituting the expression for $(\hat{\psi} - \psi)$ in equation (5.10), we have up to order $o(m^{-1})$

$$[\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi)] = L_t^T(\psi) [I_\psi^{-1}(\psi) \otimes I_m] [S_\psi(\psi) \otimes e_t] \quad (5.12)$$

and

$$\begin{aligned} & [(\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi)) (\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi))^T] \\ &= L^T(\psi) [I_\psi^{-1}(\psi) \otimes I_m] \text{Col Concat}_{1 \leq d \leq 3, 1 \leq e \leq 3} \\ & [e_t S_d(\psi) S_e(\psi) e_t^T] [I_\psi^{-1}(\psi) \otimes I_m] L(\psi) \end{aligned} \quad (5.13)$$

where

$$S_\psi(\psi) = \text{Col}_{1 \leq d \leq 3} [S_d(\psi)], \quad S_d(\psi) = \frac{\partial \ell}{\partial \psi_d}.$$

Using the expression for derivatives of likelihood, we have

$$\begin{aligned} S_d(\psi) &= \frac{1}{2} \left[\text{Trace}[C_{1d}(\psi)] - \sum_{t=1}^T \text{Trace} \left[H_t^{-1} \frac{\partial H_t}{\partial \psi_d} \right] \right. \\ &\quad \left. + \sum_{t=1}^T [e_t^T B_{1d}(\psi) e_t] \right] \\ - \left[e_t^T H_t^{-1} \frac{\partial e_t}{\partial \psi_d} \right] C_{1d}(\psi) &= \left[(X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} \frac{\partial H_1}{\partial \psi_d} H_1^{-1} X_1 \right], \\ B_{1d}(\psi) &= H_1^{-1} \frac{\partial H_1}{\partial \psi_d} H_1^{-1}. \end{aligned}$$

By applying the considerations $e_t \sim N(0, H_t)$, $\text{Corr}(e_i, e_j) = 0$ for $i \neq j$, $\text{Corr}(e_i, (\partial e_i) / (\partial \psi_d)) = 0$ and $\text{Corr}(e_i, (\partial^2 e_i) / (\partial \psi_d \partial \psi_e)) = 0$ due to the fact that $(\partial e_i) / (\partial \psi_d) = (\partial(y_i - U_i \hat{\alpha}_{it-1})) / (\partial \psi_d)$ being linear function of $(y_1, y_2, \dots, y_{t-1})$ is uncorrelated with e_t , we get the expectation of the inner most terms of the expression (5.13) as

$$\begin{aligned} E[e_t S_d(\psi) S_e(\psi) e_t^T] &= K_{de}(\psi) H_t + 2 \left[\frac{\partial H_t}{\partial \psi_d} H_t^{-1} \frac{\partial H_t}{\partial \psi_e} \right] \\ &\quad + \frac{1}{2} \left[\text{Trace}[B_{1d}(\psi)] \frac{\partial H_t}{\partial \psi_e} + \text{Trace}[B_{1e}(\psi)] \frac{\partial H_t}{\partial \psi_d} \right] \\ &\quad + \frac{1}{4} \text{Trace}[B_{1d}(\psi)] \text{Trace}[B_{1e}(\psi)] H_t \end{aligned}$$

where

$$K_{de}(\psi) = \frac{1}{2} \sum_{t=1}^T \text{Trace} \left[H_t^{-1} \frac{\partial H_t}{\partial \psi_d} H_t^{-1} \frac{\partial H_t}{\partial \psi_e} \right].$$

The middle three terms in the expression being of order $O(1)$ which along with $I_\psi^{-1}(\psi)$ in the expression given below makes them of order $o(m^{-1})$,

$$\begin{aligned} E[(\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi)) (\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi))^T] &= g_{3t}(\psi) \\ &= L^T(\psi) [I_\psi^{-1}(\psi) \otimes I_m] [K_\psi(\psi) \otimes H_t] \\ &\quad [I_\psi^{-1}(\psi) \otimes I_m] L(\psi) + o(m^{-1}) \\ &= L^T(\psi) [I_\psi^{-1}(\psi) K_\psi(\psi) I_\psi^{-1}(\psi) \otimes H_t] L(\psi) + o(m^{-1}). \end{aligned}$$

Theorem A.4: Under Regularity Conditions 2

$$\begin{aligned} E[g_{12t}(\hat{\psi}) + g_{3t}(\hat{\psi}) + g_{31t}(\hat{\psi}) - g_{4t}(\hat{\psi}) - g_{5t}(\hat{\psi})] \\ = g_{12t}(\psi) + o(m^{-1}) \\ E[g_{3t}(\hat{\psi})] = g_{3t}(\psi) + o(m^{-1}) \end{aligned}$$

and

$$E[g_{5t}(\hat{\psi})] = g_{5t}(\psi) + o(m^{-1}).$$

Proof of Theorem A.4

The proof is essentially based on the line suggested in proving Theorem A.2. Using Taylor series expansion of $g_{12t}(\hat{\psi})$ around ψ , we get

$$\begin{aligned} g_{12t}(\hat{\psi}) &= g_{12t}(\psi) + [(\hat{\psi} - \psi) \otimes I_m]^T \nabla g_{12t}(\psi) \\ &\quad + \frac{1}{2} [(\hat{\psi} - \psi)^T \otimes I_m] \nabla^2 g_{12t}(\psi) [(\hat{\psi} - \psi) \otimes I_m] \\ &\quad + o_p(m^{-1}) \end{aligned}$$

$$\nabla g_{12t}(\psi) = \text{Col}_{1 \leq d \leq 3} [\nabla g_{12td}(\psi)], \quad \nabla g_{12td}(\psi) = \frac{\partial g_{12t}(\psi)}{\partial \psi_d}$$

$$\begin{aligned} \nabla^2 g_{12t}(\psi) &= \text{Col}_{1 \leq d \leq 3} \left[\text{Concat}_{1 \leq e \leq 3} \frac{\partial^2 g_{12t}(\psi)}{\partial \psi_d \partial \psi_e} \right] \\ \frac{\partial g_{12t}(\psi)}{\partial \psi_d} &= R \Sigma^{-1} \frac{\partial R}{\partial \psi_d} \Sigma^{-1} R \\ \frac{\partial^2 g_{12t}(\psi)}{\partial \psi_d \partial \psi_e} &= -2 R \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_d} \Sigma^{-1} \frac{\partial \Sigma}{\partial \psi_e} \Sigma^{-1} R \\ &\quad + R \Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \psi_d \partial \psi_e} \Sigma^{-1} R. \end{aligned}$$

Using the fact that $\Sigma(\psi)$ and its derivatives are symmetric, we have the second term of the expression as

$$\begin{aligned} & [(\hat{\psi} - \psi)^T \otimes I_m] \nabla^2 g_{12t}(\psi) [(\hat{\psi} - \psi) \otimes I_m] \\ = & -L^T(\psi) [I_\psi^{-1}(\psi) \otimes \Sigma] L(\psi) \\ & + \frac{1}{2} \text{Trace}_m \left[[I_3 \otimes (R\Sigma^{-1})] \frac{\partial^2 \Sigma}{\partial \psi \partial \psi^T} [I_\psi^{-1}(\psi) \otimes (\Sigma^{-1}R)] \right] \\ = & -g_{3t}(\psi) = g_{5t}(\psi) \end{aligned}$$

where $I_\psi^{-1}(\psi) = \text{Var}(\psi)$ is the asymptotic variance of ψ . The first term in the expression $[(\hat{\psi} - \psi)^T \otimes I_m] \nabla g_{12t}(\psi)$ reduces to $g_{4t}(\psi)$ because of $E(\hat{\psi} - \psi) = b_\psi(\psi)$ up to order $o(m^{-1})$ (Peers and Iqbal 1985).

The second part of the Theorem follows from the Taylor series expansion of $g_{3t}(\hat{\psi})$ and $g_{5t}(\hat{\psi})$, each around ψ and using $\hat{\psi} - \psi = O_p(m^{-1/2})$ and $(\partial^2 g_{3t}(\psi)) / (\partial \psi_d \partial \psi_e) |_{\psi=\hat{\psi}} = O_p(m^{-1})$ and $(\partial^2 g_{5t}(\psi)) / (\partial \psi_d \partial \psi_e) |_{\psi=\hat{\psi}} = O_p(m^{-1})$, respectively where $\|\hat{\psi}^* - \hat{\psi}\| \leq \|\hat{\psi} - \psi\|$.

References

- Cliff, A.D., and Ord, J.K. (1981). *Spatial Processes, Models and Applications*. Pion Literature, London.
- Cox, D.R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49, 1-39 (with discussion).
- Cressie, N. (1990). Small-Area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, October 1990.
- Cressie, N. (1992). REML estimation in empirical bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.
- Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimates for best linear unbiased predictors in small area estimation problem. *Statistica Sinica*, 10, 613-627.
- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistics Association*, 74, 267-277.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
- Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistics Association*, 79, 853-862.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1997). Report No. 404(50/1.0/1), Consumption of some important commodities in India. *NSS 50th Round, July 1993-June 1994*.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Report No. 436(51/1.0/1), Household consumption expenditure and employment situation in India. *NSS 51st Round, July 1994-June 1995*.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Report No. 440(52/1.0/1), Household consumption expenditure and employment situation in India, *NSS 52nd Round, July 1995-June 1996*.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Report No. 442(53/1.0/1), Household consumption expenditure and employment situation in India. *NSS 53rd Round, January-December 1997*.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1999). Report No. 448(54/1.0/1), Household consumption expenditure and employment situation in India. *NSS 54th Round, January-June 1998*.
- National Sample Survey Organisation, Ministry of Statistics and Programme Implementation, Govt. of India (2001). Report No. 472(55/1.0/1), Differences in level of consumption among socio-economic groups. *NSS 55th Round, July 1999-June 2000*.
- Peers, H.W., and Iqbal, M. (1985). Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications. *Journal of the Royal Statistical Society, Series B*, 47, 547-554.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean square error of small area estimates. *Journal of the American Statistics Association*, 85, 163-171.
- Rao, C.R., and Toutenbourg, Heldge (1999). *Linear Models: Least Squares and Alternatives*. Second Edition, New York: Springer.
- Rao, J.N.K. (1999). Some recent advances in model based small area estimation. *Survey Methodology*, 25, 175-186.
- Ripley, B.D. (1981). *Spatial Statistics*. New York: John Wiley & Sons, Inc.
- Sallas W.M., and Harville D.A. (1994). Noninformative priors and restricted likelihood estimation in the Kalman filter. *Bayesian Analysis of Time Series and Dynamic Models*, (Ed. James C. Spall). New York: Marcel Dekker Inc., 477-508.
- Singh, A.C., Mantel, H.J. and Thomas B.W. (1994). Time series EBLUPs for small area using survey data. *Survey Methodology*, 20, 33-43.
- Singh, A.C., Stukel, D. and Pfeiffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, 377-396.
- Srivastava, V.K., and Tiwari, R. (1976). Evaluation of expectations of products of stochastic matrices. *Scandinavian Journal of Statistics*, 3, 135-138.

Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey

Liv Belsby, Jan Bjørnstad and Li-Chun Zhang¹

Abstract

This paper considers the problem of estimating, in the presence of considerable nonignorable nonresponse, the number of private households of various sizes and the total number of households in Norway. The approach is model-based with a population model for household size given registered family size. We account for possible nonresponse biases by modeling the response mechanism conditional on household size. Various models are evaluated together with a maximum likelihood estimator and imputation-based poststratification. Comparisons are made with pure poststratification using registered family size as stratifier and estimation methods used in official statistics for The Norwegian Consumer Expenditure Survey. The study indicates that a modeling approach, including response modeling, poststratification and imputation are important ingredients for a satisfactory approach.

Key Words: Household size; Nonresponse; Imputation; Poststratification.

1. Introduction

This work is motivated by the considerable nonresponse rate in the Norwegian Consumer Expenditure Surveys (CES) for private households, for example 32% in the 1992 survey. Nonresponse involves both noncontact and refusal. We focus on the problem of nonignorable nonresponse that occurs when estimating the number of households of various sizes and the total number of households.

We shall consider a completely model-based approach; modeling and estimating the distribution of household size given registered family size and the response mechanism conditional on the household size. This model takes into account that the nonresponse mechanism may be nonignorable, in the sense that the probability of response is allowed to depend on the size of the household. The response model is used to correct for nonresponse. Model-based approaches with nonresponse included, sometimes called the prediction approach, have been considered by, among others, Little (1982), Greenlees, Reece and Zieschang (1982), Baker and Laird (1988), Bjørnstad and Walsøe (1991), Bjørnstad and Skjold (1992) and Forster and Smith (1998).

For various models of household size and response we consider mainly two model-based approaches, a maximum likelihood estimator and imputation-based poststratification after registered family size. These methods are compared to pure poststratification and the methods in current use in CES.

The main issue here is a comparison of models and methods with estimation bias as the basic problem. In addition, standard errors of the estimates and differences of the estimates, conditional on the sizes of post-strata determined by family size, are estimated using a bootstrap approach. In addition to assessing the statistical uncertainty of the estimators, this is done to help evaluate the extent to which differences between the proposed estimators are attributable to sampling error, nonresponse bias or both. However, in this evaluation we keep in mind the following quote from Little and Rubin (1987, page 67): "It is important to emphasize that in many applications the issue of nonresponse bias is often more crucial than that of variance. In fact, it has been argued that providing a valid estimate of sampling variance is worse than providing no estimate if the estimator has a large bias, which dominates the mean squared error."

Section 2 describes the data-structure and the sample design of CES, and Section 3 considers modeling issues. Section 3.1 presents the various models for household size and response to be considered for the 1992 CES, Section 3.2 describes the maximum likelihood method for parameter estimation, and in Section 3.3 the models are evaluated. A family size group model for household size and a logistic link for the response probability using household size as a categorical variable give the best fit of the models under consideration. Section 3.4 gives the estimated household size distributions for different family sizes and estimated response probabilities for different household sizes.

1. Liv Belsby, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: lbe@ssb.no; Jan F. Bjørnstad, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: jab@ssb.no and Li-Chun Zhang, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: lc2@ssb.no.

Section 4 considers model-based estimation, the imputation method, imputation-based estimators and the variance estimation method. It is shown that for the chosen model for household size from Section 3.3, the maximum likelihood estimator and the imputation-based poststratified estimator are identical.

Section 5 deals with the main goal of estimating the total number of household of various sizes based on the 1992 CES, using the estimators in Section 4. The model that gave the best fit seems to work well for our estimation problem. We conclude that poststratification, response modeling and imputation are key ingredients for a satisfactory approach.

2. Norwegian Consumer Expenditure Survey

The population totals within household-size categories provide a more correct number of dwellings than the totals within family-size categories from the Norwegian Family Register. Furthermore, the authorities for evaluating eventual policy intervention aimed at housing construction use the estimated number of households. Estimating household-size totals is therefore an important issue in social planning. It is invariably affected by nonignorable nonresponse, no matter what kind of survey one uses. Hence, it is a good illustration for how to handle nonresponse bias. We shall base our estimation on the Norwegian Consumer Expenditure Surveys (CES), where it is important to gain information about the composition of households, since household size influences consumption.

The actual CES, the survey for expenditure variables, is a sample of private households from all private households in Norway. This is done by selecting a sample of persons and including the whole households these persons belong to. Persons older than 80 years old are excluded since they often live in institutions. For our purpose, the units of interest in the survey are *persons* between the ages of 16 and 80 living in private households, and the variable of interest is the size of the *household* the person belongs to, which is observed only in the response sample of the persons selected.

The sample design is a three-stage self-weighting sample of persons. That is, every person in the population has the same inclusion probability to the total sample. The first two stages select geographical areas in a stratified way, while at the third stage persons are selected randomly from the chosen geographical areas. The primary sampling units (PSU) at stage 1 consists of the municipalities in Norway. Municipalities with less than 3,000 inhabitants are grouped together such that each PSU consists of at least 3,000 persons. The PSUs are first grouped into 10 regions and within each region stratified according to size (number of inhabitants) and type of municipality (*i.e.*, industrial

structure and centrality). Totally, we have 102 strata. Towns of more than 30,000 inhabitants are their own strata and therefore selected with certainty at stage 1. For the other strata, one PSU is selected with probability proportional to size. At the second stage, the selected PSUs are divided into three smaller areas (secondary sampling units, SSU) and one of these is selected at random. Finally, at the third stage, for each of the selected SSU, a random sample of persons is selected. The sample sizes for each selected SSU are determined such that the resulting total sample of persons is self-weighting.

Our application is based on the data from the 1992 CES. CES is a yearly survey and since 1992 a modified Horvitz-Thompson estimator, including a correction for nonresponse by estimating response probabilities given household size, has been employed (see Belsby 1995). The weights equal the inverse of the probability of being selected multiplied with the conditional probability of response given selected. Since 1993 the probability of response is estimated with a logistic model with auxiliary variables being place of residence (rural/urban), and household size. For most of the nonrespondents the family size is used as a substitute for the household size.

A household is defined as persons having a common dwelling and sharing at least one meal each day (having common board). For a complete description of CES we refer to Statistics Norway (1996). In CES, the auxiliary variables known for the total sample, including the nonrespondents, are the family size, the time of the survey (summer/not summer), and the place of residence (urban/rural). *Families* are registered in Norwegian Family Register, (*NFR*), and may differ from the household the persons in the family belong to, both by definition and because of changes not yet registered. Hence, the registered *family* size from *NFR* differs to some extent from the household size. Initially, based on experience from previous surveys, all the auxiliary variables and household size are assumed to affect the response rate.

Table 1 shows the data for the 1992 CES with a total sample of 1,698 persons. The households with size five and greater are collapsed due to the low frequency in the sample of households. We base our modeling and estimation on two corresponding tables, one for the persons in rural areas and one for the persons in urban areas. These data are given in table A1 in appendix A1.

For example, the number 48 in cell (1,2) means that of the 162 persons registered to live alone in the response sample, 48 are actually living in a two-persons household. This is explained mostly by young people's tendency to cohabitate without being married; see Keilman and Brunborg (1995).

Table 1
Family and household sizes for the 1992 Norwegian Consumer Expenditure Survey

Family size	Household size					Total	Nonresponse	Response rate
	1	2	3	4	≥ 5			
1	83	48	20	9	2	162	153	0.514
2	9	177	37	4	3	230	160	0.590
3	10	25	131	40	6	212	91	0.700
4	2	13	37	231	17	300	123	0.709
≥ 5	1	4	4	17	181	207	60	0.775
Total	105	267	229	301	209	1,111	587	0.654

3. Modeling of Household Size and Nonresponse

We shall assume a population model for the household size, given auxiliary variables, *i.e.*, we model the conditional probability. To take nonresponse into account in the statistical analysis, we must model the response mechanism, *i.e.*, the distribution of response conditional on the household size and auxiliary variables. The sampling mechanism for persons is ignorable for the survey we consider, *i.e.*, is independent of the population vector of household sizes. The statistical analysis is therefore done *conditional* on the total sample, following the likelihood principle (see Bjørnstad 1996). Hence, probability considerations based on the sampling design is irrelevant in the statistical analysis. This is the so-called prediction approach. However, when evaluating the estimation methods with regard to statistical uncertainty, we do this from a common randomization perspective as described in Section 4.3.

For CES, the auxiliary vector consists of the family size, place of residence divided into rural and urban areas, and time of the data collection.

3.1 The Models

Let us first consider a simple model for the household size, denoted by Y . Let \mathbf{x} denote all auxiliary variables. The household size is assumed to depend only on the family size x , and as such is a model with a restricted parametric link function, but with no additional assumptions,

$$P(Y_i = y | \mathbf{x}_i) = P(Y_i = y | x_i) = p_{y, x_i}, \quad (3.1)$$

where

$$\sum_y p_{y, x_i} = 1, \text{ for each possible value of } x_i.$$

The model (3.1) is flexible in the sense that it does not include any restrictions on the assumed model function of x_i . The drawback is the high number of parameters compared with a model using a logistic type model with a linear, in \mathbf{x} , link function (the function linking $P(Y = y)$ with \mathbf{x}). If nonresponse is ignored the estimates in this model would simply be the observed rates.

Household size defines ordered categories. Thus a natural choice for a model is the cumulative logit model, known as the proportional-odds model (see McCullagh and Nelder 1991), assuming (with θ_y increasing in y)

$$P(Y_i \leq y | \mathbf{x}) = \begin{cases} \frac{1}{1 + \exp(-\theta_y + \beta' \mathbf{x})} & \text{for } y = 1, 2, 3, 4 \\ 1 & \text{for } y \geq 5. \end{cases}$$

However, a goodness of fit test, with \mathbf{x} consisting of family size and place of residence, indicated that this model fits the data badly. Thus we choose to reject it.

It is assumed that the probability of nonresponse may depend on the household size. For example, one-person households are less likely to respond than households of larger size since larger households are easier to “find at home”. Nonresponse is indicated by the variable R , where $R_i = 1$ if person i responds and 0 otherwise. Let R_s be the vector of these indicators in the total sample. From Bjørnstad (1996), the response mechanism (RM), *i.e.*, the conditional distribution of R_s given the \mathbf{x} -values in the population and the y -values in the total sample, is defined to be ignorable if it can be discarded in a likelihood-based analysis. This means that RM is ignorable if this conditional distribution of R_s does not depend on the unobserved y -values, coinciding with the definition used by Little and Rubin (1987, pages 90, 218). For our case it is assumed that all pairs (Y_i, R_i) are independent. Then RM is ignorable if Y_i and R_i are independent. Hence, nonignorable response mechanism is equivalent to

$$P(Y_i = y_i | \mathbf{x}_i, r_i = 0) \neq P(Y_i = y_i | \mathbf{x}_i, r_i = 1)$$

and then both are different from $P(Y_i = y_i | \mathbf{x}_i)$.

Thus estimating the parameters in the model for $P(Y = y | \mathbf{x})$ using only the response sample, ignoring that the probability of response depends on the household size, would most likely give biased estimates for the unknown parameters. Also the poststratification estimator would give

biased estimates because it assumes that the distribution of R only depends on the auxiliary \mathbf{x} . *E.g.*, the observed lower response rate among one-person families indicates that the same may hold for one-person households. If so, the estimated probability of household size 1, based on respondents only, would be too small. Poststratification with respect to family size will most likely correct only some of this bias.

The model for the probability of response, given auxiliary variables and household size y_i , is assumed to be logistic. It depends on the auxiliary variables \mathbf{z}_i , which includes part of \mathbf{x}_i , expressed by

RM1(y, \mathbf{z}):

$$P(R_i = 1 | y_i, \mathbf{z}_i) = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi' \mathbf{z}_i)}. \quad (3.2)$$

Here, α and γ are scalar parameters and ψ is a vector. The variable y_i has an order. Motivated by this fact, and to avoid introducing many parameters, y_i is used in (3.2) as an ordinal variable rather than a class variable. Thus the logit function,

$$\log\{P(R_i = 1 | y_i, \mathbf{z}_i) / P(R_i = 0 | y_i, \mathbf{z}_i)\} = \alpha + \gamma y_i + \psi' \mathbf{z}_i,$$

is linear in y_i . To avoid the assumption of linear logit in y_i , we also consider a model with y_i as a categorical variable, *i.e.*,

$$\text{RM2}(y, \mathbf{z}): P(R_i = 1 | y_i, \mathbf{z}_i) = \frac{1}{1 + \exp\left(\begin{matrix} -\alpha_0 - \alpha_1 I_1(y_i) - \alpha_2 I_2(y_i) \\ -\alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi' \mathbf{z}_i \end{matrix}\right)}, \quad (3.3)$$

where the indicator variable $I_y(y_i)$ equals 1 if $y_i = y$ and 0 otherwise. The drawback with this model is that it includes three parameters more than model (3.2).

3.2 Maximum Likelihood Parameter Estimation

All the selected persons in the sample are from different households (duplicates have been removed). The population model then assumes that the household sizes Y_i are statistically independent. For *this* variable, interviewer- or cluster- effect plays no role.

Let us consider the likelihood function for estimating the unknown parameters, assuming that all pairs (Y_i, R_i) are independent and response model RM1 given by (3.2). To simplify notation we relabel the observations such that observations 1 to n_r are the respondents and observations $n_r + 1$ to n are the nonrespondents. With response model RM2 the expression for the likelihood is of the same form with (3.3) replacing (3.2).

For the respondents let $L_i = P(Y_i = y_i \cap R_i = 1 | \mathbf{x}_i)$. Then, for model (3.1)

$$L_i = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi' \mathbf{z}_i)} \cdot p_{y_i, \mathbf{x}_i}, \quad i = 1, \dots, n_r \quad (3.4)$$

For the nonrespondents let $L_i = P(R_i = 0 | \mathbf{x}_i)$. Then

$$L_i = \sum_{j=1}^5 \frac{1}{1 + \exp(\alpha + \gamma y_j + \psi' \mathbf{z}_i)} \cdot p_{y_j, \mathbf{x}_i}, \quad i = n_r + 1, \dots, n. \quad (3.5)$$

The likelihood function for the entire sample of persons from different households is given by

$$L(\theta, \beta, \alpha, \gamma, \psi) = \prod_{i=1}^n L_i. \quad (3.6)$$

For $i = 1, \dots, n_r$, L_i is according to (3.4) and for $i = n_r + 1, \dots, n$, L_i is given by (3.5).

Estimates are found by maximizing the likelihood function (3.6). The maximization was done numerically using the software TSP (1991) see Hall, Cummins and Schnake (1991). The optimizing algorithm is a standard gradient method, using the analytical first and second derivatives. These are obtained by the program, saving us a substantial piece of programming. The model fitting is based on the chi-square statistic and on the t -values, provided by TSP, where the standard errors are derived from the analytical second derivatives. The t -values have to be interpreted with some care, since the unbiasedness of the estimated standard errors depends on how well the model is specified as well as the number of observations compared with the number of parameters.

3.3 Evaluation of the Models for Household Size and Response

We present the fit of the models with the Pearson goodness-of-fit statistics. The model study is based on the 1992 CES. The parameters are considered to be significant when the absolute t -values are greater than 2. However, we do not want a model that is too restrictive, and therefore some variables are kept even though their absolute t -values are less than 2.

In the response models RM1 and RM2 we use the variable $\mathbf{z} = z$, place of residence. We let $z = 0$ if rural area and $z = 1$ if urban area. It was observed in the CES 1986–88 and CES 1992–94, see Statistics Norway (1990, 1996), that there is more nonresponse during the summer. Therefore, the time of the survey was also included in the model, that is whether or not the data were collected in the period May 21–August 12. However, the time of the survey was found to be nonsignificant, with t -value clearly less than 2. Also the family size was found to be nonsignificant. But if the household size is omitted in the response model then the family size turns out to be significant.

Ideally, we want to take a look at the empirical logit function for response with respect to the household size. However, household size is unavailable for the non-respondents. As a replacement we plot the logit-function against the family size; see figure 1. From family size one to two the two functions for rural and urban families increase in a fairly parallel way. However, for family size three and four the logit functions depart from being linear and parallel. Thus we suspect that coding the household size as a categorical variable, as in model RM2, will give better fit than restricting the logit functions to be parallel for rural and urban and linear with respect to the household size, as in model RM1.

In order to test the goodness of fit of the models, we consider the Pearson chi-square statistic, conditional on the auxiliary variables x, z . Given rural or urban type of residence and registered family size, there are six possible outcomes; household sizes 1, ..., 5 and nonresponse. Altogether there are ten multinomial trials and sixty cells. For family sizes (1,2) and (4,5), the extreme household sizes (4,5) and (1,2), respectively, are combined because the expected sizes under the models are too small. This reduces the number of cells to 52. The degrees of freedom (d.f.) is

calculated as: number of cells – number of trials – number of parameters. For model (3.1) & RM1(y, z), d.f. = $52 - 10 - (20 + 3) = 19$, and for (3.1) & RM2(y, z), d.f. = $52 - 10 - (20 + 6) = 16$. For model (3.1) & RM1(y, z) the Pearson statistic χ^2 is 26.35 and the p -value is 0.121. And for model (3.1) & RM2(y, z) χ^2 is 21.77 and the p -value is 0.151.

By studying the standardized residuals, $(\text{observed} - \text{expected}) / \sqrt{\widehat{\text{Var}}(\text{observed})}$, we find that the main reason for the better fit is that model (3.1) & RM2(y, z) does a better job of predicting the observed counts for the urban area where the response rate is lowest (see appendix A1). Thus the data indicates that coding the household size as a categorical variable, as in RM2, improves the fit compared to using it as an ordinal variable. The model (3.1), with the restricted parametric link function, combined with RM2 is the best of the models we have considered so far.

3.4 Estimated Household Size Distribution and Response Probabilities

Table 2 displays the estimates for the population model (3.1) together with the logistic response model RM2 in (3.3).

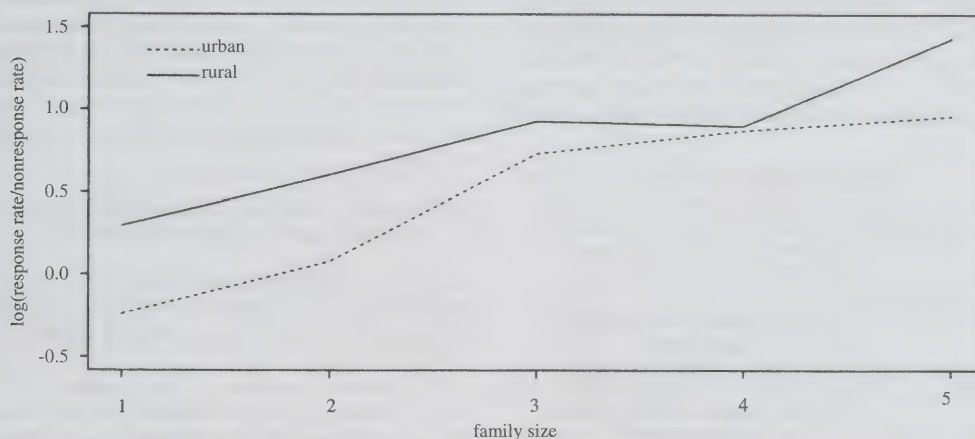


Figure 1. The logit function for the empirical response rate with respect to family size 1, ..., 5 in urban and rural areas, respectively. The computation is based on respondents and nonrespondents from Table 1 in Appendix A1.

Table 2

1992 CES. Parameter Estimates, in Percentages, for the Population Model with a Restricted Parametric Link Function, $p_{y,x}$, Combined with the Logistic Response Model RM2(y, z). In Parentheses are the Estimates for the Population Model, Ignoring the Response Mechanism

Family size, x	Household size				
	1	2	3	4	5 or more
1	60.01 (51.23)	26.75 (29.63)	8.35 (12.35)	4.09 (5.56)	0.80 (1.23)
2	5.27 (3.91)	79.80 (76.98)	12.48 (16.09)	1.47 (1.74)	0.98 (1.30)
3	7.53 (4.72)	14.45 (11.79)	56.67 (61.79)	18.85 (18.87)	2.50 (2.83)
4	1.06 (0.67)	5.31 (4.33)	11.38 (12.33)	77.20 (77.00)	5.05 (5.67)
5 or more	0.84 (0.48)	2.60 (1.93)	1.96 (1.93)	9.05 (8.21)	85.55 (87.44)

Let us interpret some of the values in the household model. Taking the response mechanism into account has largest effect on the estimated household distribution for one-person families. The probability that a household size equals one, given that the family size is one, is estimated as 60.01%. The estimate based on the traditional approach, ignoring the nonresponse, is 51.23%. The response model “adjusts” the observed rate among the respondents to a higher value. This seems reasonable since the rate of non-respondents is higher for small households. The estimated probability of household size five or more, given family size of five or more is 85.55%, which differs little from the observed rate among the respondents, 87.44%. This indicates that, given family size five or more, the household size distribution is about the same among respondents and nonrespondents.

Table 3 presents the estimated response probabilities based on RM2 in combination with the population model (3.1). Furthermore, we present estimated response probabilities based on a saturated model, with perfect fit, presented in Section 4.2. The model, defined by (4.9), assumes that the response probability for persons with the same household size within rural/urban area, respectively, is identical for different family sizes. Moreover, the model for household size depends on place of residence and family size, but with no restriction on the link function. We note that $RM2(y, z)$ satisfies (4.9b), but is more restrictive. Model (4.9) allows for more freedom than model (3.1) with $RM2(y, z)$.

Table 3

Estimated Probability of Response Based on the Logistic Model RM2 in Combination with (3.1), and the Saturated Model (4.9). The Estimates are Given in Percentages

Place of residence	Household size				
	1	2	3	4	5 or more
Estimated response probabilities for model RM2					
Rural	47.77	60.90	79.16	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46
Estimated response probabilities for the saturated model					
Rural	50.79	62.37	76.90	70.57	83.07
Urban	35.17	50.85	74.79	70.68	72.89

The estimated response probabilities reflect the lower response rate among one-person households, and the lower response rate in urban areas. Households of size five and higher have the highest response rate. The models estimate, surprisingly maybe, that the the probability of response is higher for households of size three than for households of size four. This may be explained by the fact that women often choose to have two children, and that three-person-households mostly consist of mother, father and a *small* child. Such a family will tend to stay at home and thus be

more accessible than a typical four-persons-family with two older children.

The higher estimated response rate for households of size three compared to size four is equivalent to the ratio $P(Y = 3 | R = 1) / P(Y = 3 | R = 0)$ being greater than the ratio $P(Y = 4 | R = 1) / P(Y = 4 | R = 0)$. This is consistent with the household distribution in table 2, where we estimate that $P(Y = 4) \approx P(Y = 4 | R = 1)$, i.e., $P(Y = 4 | R = 0) \approx P(Y = 4 | R = 1)$. On the other hand, the estimates in table 2 indicate that $P(Y = 3 | R = 1) > P(Y = 3)$ which means that $P(Y = 3 | R = 1) > P(Y = 3 | R = 0)$.

We see that the logistic model RM2 combined with the population model with the restricted parametric link $p_{y,x}$ acts as a smoother of the estimates based on the saturated model in (4.9), because of the added assumption of parallel logits of the response probabilities for urban and rural areas.

4. Estimators for Household Size Totals

In this section we present the estimators for household size totals and the method for variance estimation. We use a maximum likelihood estimator with the restricted parametric link function in (3.1) as population model. It is shown that this estimator is identical to an imputation-based poststratified estimator, which again turns out as a standard poststratification when the response mechanism is ignored. Furthermore, we present an imputed poststratified estimator, based on a saturated model for household size and response probability.

4.1 Estimators Based on a Restricted Parametric Link Function as Population Model

With N_y denoting the total number of persons living in households of size y , the number of households of size y equals $H_y = N_y / y$. The total number of households is denoted by H , $H = \sum_y H_y$.

The statistical problem is to estimate H_y for $y = 1, \dots, J$ and H . The largest size J is chosen such that there are few households of size greater than J . Strictly speaking, H_J is the number of households of size J or more, and likewise for N_J . In our application we choose $J = 5$ due to the low frequency in the sample of households of size greater than five. We can write $N_y = \sum_{i=1}^N I(Y_i = y)$, where the indicator function $I(Y_i = y) = 1$ if $Y_i = y$, and 0 otherwise. Hence, with $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$,

$$E(H_y | \mathbf{x}) = \frac{1}{y} \sum_{i=1}^N P(Y_i = y | \mathbf{x}_i).$$

A maximum likelihood based estimator for H_y can be obtained by estimating $E(H_y | \mathbf{x})$, i.e., replacing $P(Y_i = y | \mathbf{x}_i)$ by the maximum likelihood estimator

$\hat{P}(Y_i = y | x_i)$. The data is stratified according to family sizes 1, ..., K , where the last category contains persons belonging to families of sizes $\geq K$. Using the model with the restricted parametric link function, defined in (3.1), Y is assumed to depend only on the family size x , and the estimator takes the form

$$\hat{H}_y = \frac{1}{y} \sum_{x=1}^K M_x \hat{P}(Y = y | x) \quad (4.1)$$

where M_x (M_K) denotes the number of persons in the population with registered family size x ($\geq K$). The M_x 's are known auxiliary information from the Norwegian Family Register.

A common approach to correct for nonresponse is by imputation of the missing values in the sample. Based on the estimated distribution for Y for a given family size and place of residence for the nonrespondents, $\hat{P}(Y = y | x, z, r = 0)$, we assign the nonrespondents to the values 1, ..., 5 in proportions given by $\hat{P}(Y = y | x, z, r = 0)$ for $y = 1, \dots, 5$. Let $n_{xy}^*(0)$ ($n_{xy}^*(1)$) be the number of imputed values with family size x and household size y , for rural (urban) areas and let $m_{xu}(0)$ ($m_{xu}(1)$) be the number of missing observations for persons in rural (urban) areas with family size x . Then

$$n_{xy}^*(z) = m_{xu}(z) \cdot \hat{P}(Y = y | x, z, r = 0), z = 0, 1. \quad (4.2)$$

and

$$n_{xy}^* = n_{xy}^*(0) + n_{xy}^*(1)$$

is the total number of imputed values with family size x and household size y , i.e., n_{xy}^* is the estimated expected number of households of size y , given family size x and $r = 0$.

The following general result holds, showing that with population model (3.1), the maximum likelihood estimator (4.1) is identical to an imputation-based poststratified estimator.

Theorem. Assume model (3.1) for Y . That is, $P(Y = y | x, z) = p_{y,x}$ is independent of z , but otherwise the $p_{y,x}$'s are completely unknown with the only restriction $\sum_y p_{y,x} = 1$, for all values of x . The response mechanism is arbitrarily parametrized, i.e., no assumption is made about $P(R = 1 | Y = y, x, z)$. Then the maximum likelihood estimates for $p_{y,x}$ are given by, for $x = 1, \dots, K$,

$$\hat{p}_{y,x} = \frac{n_{xy} + n_{xy}^*}{m_x + m_{xu}},$$

where n_{xy} is the number of respondents belonging to a family of size x and household size y , m_x (m_K) is the number of respondents belonging to families of size x ($\geq K$), and $m_{xu} = m_{xu}(0) + m_{xu}(1)$.

Proof. See Appendix A2.

The theorem implies that the estimator can be written as the imputation-based poststratified estimator, using family size as the stratifying variable,

$$\hat{H}_{y, \text{post}}^I = \frac{1}{y} \sum_{x=1}^K M_x \frac{n_{xy} + n_{xy}^*}{m_x + m_{xu}}. \quad (4.3)$$

Assuming ignorable response mechanism and using the model (3.1), the likelihood function is given by $\prod_{i=1}^{n_r} P(Y_i = y_i | x_i)$. Then the maximum likelihood estimate $\hat{P}(Y = y | x)$ is simply the observed rate among the respondents with household size y , given family size x . Thus the maximum likelihood estimator turns out to be identical to the standard poststratified estimator, with family size as the stratifying variable,

$$\hat{H}_{y, \text{post}} = \frac{1}{y} \sum_{x=1}^K M_x \frac{n_{xy}}{m_x}. \quad (4.4)$$

For a general study of poststratification see, for example Holt and Smith (1979) and Särndal, Swensson and Wretman (1992, chapter 7.6).

To illustrate the effects of nonresponse modeling and poststratification, we also present estimates based on the regular expansion estimator, given by

$$\hat{H}_{y, e} = \frac{1}{y} \cdot N \frac{n_y}{n_r} \quad (4.5)$$

and the imputation-based expansion estimator given by

$$\hat{H}_{y, e}^I = \frac{1}{y} \cdot N \frac{n_y + n_y^*}{n}. \quad (4.6)$$

Here, n_y is the number of respondents in households of size y , n_r is the total number of respondents, and $n_y^* = \sum_x n_{xy}^*$. The estimator (4.5) does not seek to correct for nonresponse nor use the family population distribution as a post-stratifying tool to improve the estimation, while estimator (4.6) tries to take the response mechanism into account, but cannot correct for nonrepresentative samples.

4.2 Imputation-based Poststratification with a Saturated Model

We now proceed to an intuitive method of imputation that was used to estimate response probabilities for a modified Horvitz-Thompson estimator in the official statistics from the 1992 CES (described in Belsby 1995). We will use this imputation method for the poststratified estimator (4.3).

The imputation method consists of distributing, within rural/urban area, the $m_{xu}(z)$ nonresponse units over the household sizes 1, ..., 5 in such a way that, given

household size, the rate of nonresponse is the same for all family sizes. It implicitly assumes that the response probability for persons with the same household size within rural/urban area is identical for different family sizes. Denote the number of nonresponse persons with family size x and household size y and place of residence z obtained in this manner by $h_{xy}(z)$. The corresponding number among the respondents is $n_{xy}(z)$. The values of $h_{xy}(z)$ are determined by the equations

$$\frac{h_{xy}(z)}{h_{xy}(z) + n_{xy}(z)} = \frac{h_{iy}(z)}{h_{iy}(z) + n_{iy}(z)}, \quad z=0, 1. \quad (4.7)$$

When $n_{xy}(z)=0$, we let $h_{xy}(z)=0$. The equation (4.7) is solved under the conditions

$$\sum_y h_{xy}(z) = m_{xu}(z); x=1, 2, 3, 4, 5 \text{ and } z=0, 1. \quad (4.8)$$

Solving (4.7) and (4.8) requires, for each value of z , one row $(n_{x1}(z), n_{x2}(z), \dots, n_{x5}(z))$ of nonzeros, which holds for our case. The imputed values $h_{xy}(z)$ determined by (4.7) and (4.8) correspond to the imputation method described by (4.2) for the following model:

$$P(Y=y|x, z) = p_{y, x, z} \text{ with no restrictions} \quad (4.9a)$$

$$P(R=1|Y=y, x, z) = q_{y, z}, \text{ independent of } x. \quad (4.9b)$$

This can be seen as follows:

For the ten multinomial trials determined by the different (x, z) -values, we have 50 unknown cell probabilities $\pi_{y, x, z} = P(Y=y, R=1|x, z)$. With no restrictions on cell probabilities, the maximum likelihood estimates (mle) are given by observed relative frequencies,

$$\hat{\pi}_{y, x, z} = \frac{n_{xy}(z)}{m_x(z) + m_{xu}(z)}.$$

This also holds when $n_{xy}(z)=0$. Now, it can be shown that there is a one-to-one correspondence between $\pi = (\pi_0, \pi_1)$ and (p_0, q_0, p_1, q_1) , where $\pi_z = (\pi_{y, x, z} : y=1, \dots, 5; x=1, \dots, 5)$, $p_z = (p_{y, x, z} : y=1, \dots, 5; x=1, \dots, 5)$ and $q_z = (q_{1, z}, \dots, q_{5, z})$. Since $\pi_{y, x, z} = p_{y, x, z} \cdot q_{y, z}$, the mle of $p_{y, x, z}$ and $q_{y, z}$ must satisfy

$$\hat{p}_{y, x, z} \cdot \hat{q}_{y, z} = \frac{n_{xy}(z)}{m_x(z) + m_{xu}(z)} \quad (4.10)$$

and are uniquely determined by $\hat{\pi}_{y, x, z}$.

Consider $h_{xy}(z)$, given by (4.5) & (4.6). Let $h_y(z) = \sum_x h_{xy}(z)$ and $n_y(z) = \sum_x n_{xy}(z)$. From (4.7),

$$\frac{h_j(z)}{h_j(z) + n_j(z)} = \frac{h_{xj}(z)}{h_{xj}(z) + n_{xj}(z)}, \text{ if } n_{xj}(z) > 0. \quad (4.11)$$

From (4.10) and (4.11) we have that the following intuitive estimates also are mle.

$$\hat{q}_{y, z} = \frac{n_y(z)}{n_y(z) + h_y(z)} \quad (4.12)$$

$$\hat{p}_{y, x, z} = \frac{n_{xy}(z) + h_{xy}(z)}{m_x(z) + m_{xu}(z)} \quad (4.13)$$

(also when $n_{xy}(z) = h_{xy}(z) = 0$).

(We can also show (4.12) and (4.13) by maximizing the loglikelihood directly.) Next, we show that the imputed values (4.2) for the model (4.9) equal $h_{xy}(z)$. From (4.2), we have $n_{xy}^*(z) = m_{xu}(z) \cdot \hat{P}(Y=y|x, z, r=0)$. Under model (4.9) and estimates (4.12) and (4.13), we find that

$$\begin{aligned} \hat{P}(Y=y|x, z, R=0) &= \frac{\hat{P}(Y=y|x, z) - \hat{P}(Y=y, R=1|x, z)}{\hat{P}(R=0|x, z)} \\ &= \frac{\hat{p}_{y, x, z} - \hat{\pi}_{y, x, z}}{1 - \sum_y \hat{\pi}_{y, x, z}} \\ &= \frac{n_{xy}(z) + h_{xy}(z) - n_{xy}(z)}{m_{xu}(z)} = \frac{h_{xy}(z)}{m_{xu}(z)}, \end{aligned}$$

and it follows that $n_{xy}^*(z) = h_{xy}(z)$. If $n_{xy}(z) = 0$, then $\hat{p}_{y, x, z} = \hat{\pi}_{y, x, z} = 0$, and $n_{xy}^*(z) = 0$. We note that model (4.9) is saturated and will, from (4.10), give perfect fit.

The imputation-based expansion estimates (4.6), with model (4.9), are identical to the modified Horvitz-Thompson estimates with $\hat{q}_{y, z} = n_y(z)/[n_y(z) + n_y^*(z)]$ (from (4.12)) as the estimated response probabilities, used in the official statistics from the 1992 CES. This follows from the fact that the modified Horvitz-Thompson estimator of N_y is given by

$$\hat{N}_{y, HT} = \sum_{i \in s_r} \frac{I(Y_i = y)}{\pi_i},$$

where $\pi_i = P(\text{person } i \text{ is selected to the sample and responds})$. Hence,

$$\pi_i = \frac{n}{N} \hat{P}(R_i=1|x_i, z_i, Y_i=y) = \frac{n}{N} \hat{q}_{y, z_i}$$

and

$$\hat{N}_{y, HT} = \frac{N}{n} \left(\frac{n_y(0)}{\hat{q}_{y, 0}} + \frac{n_y(1)}{\hat{q}_{y, 1}} \right). \quad (4.14)$$

Here,

$$\begin{aligned} \hat{N}_{y, HT} &= \frac{N}{n} \left(\frac{n_y(0)}{n_y(0)/(n_y(0) + n_y^*(0))} + \frac{n_y(1)}{n_y(1)/(n_y(1) + n_y^*(1))} \right) \\ &= N \frac{n_y + n_y^*}{n}. \end{aligned}$$

So this modified Horvitz-Thompson estimator suffers from the same negative feature as the imputation-based expansion estimator (4.6); it cannot correct for the bias in an unrepresentative sample. For a general description of the modified Horvitz-Thompson method see, *e.g.*, Särndal *et al.* (1992, chapter 15).

4.3 Variance Estimation

Variance estimation of the various estimates are obtained by bootstrapping. It can be carried out under the modeling or quasi-randomization framework (Little and Rubin 1987). For instance, to estimate the variance under model (3.1) and RM1 (3.2), we may apply the parametric bootstrap with the estimated parameters (Efron and Tibshirani 1993). However, it is not clear how to compare the variances estimated under the alternative models. We have therefore chosen to estimate the variances of the different estimators under a common quasi-randomization framework. We assume simple random sampling conditional to the family size, which is the only assumption we make for variance estimation. Unconditionally we have a self-weighting, but not simple random, sample, and therefore this is a rather crude approximation to the actual conditional sampling design. However, for a comparative study of the estimators the approximation will serve this purpose well. The nonresponse indicator r_i is considered to be a constant associated with person i . We draw the bootstrap sample, resampling $(y_i, z_i, r_i = 1), (z_i, r_i = 0)$ randomly with replacement, as described by Shao and Sitter (1996, Section 5), within each post-stratum of $\{i; x_i = x\}$. While the sizes of the sample post-strata are fixed, both the number of nonrespondents and the number of persons from urban or rural areas vary from one bootstrap sample to another. We calculate the bootstrap estimates in the same way as based on the observed data. In particular, the bootstrap data are imputed in the same way as the original data if the estimator is imputation-based. Finally, the estimated variances and standard errors are obtained by the usual Monte Carlo approximation based on 500 independent bootstrap samples.

5. Estimated Number of Households of Different Sizes Based on the 1992 Norwegian Consumer Expenditure Survey

In this section we present the estimated number of households of sizes one to five and more, and the total number of households for the population in Norway aged less than eighty years old. The estimation uses the data from CES 1992, and is based on the estimators considered in Section 4. To compute the estimates we need the number of families of different sizes in the population, *i.e.*, M_x , at the time of the 1992 survey. The actual number at the time of

the survey is not recorded. As an approximation we use the numbers at January 1, 1993. These are given in table 4.

Table 4
Families and Persons with Age Less than 80 Years
in Norway at January, 1993

Number of persons in family	Families	Persons
1 person	793,869	793,869
2 persons	408,440	816,880
3 persons	261,527	784,581
4 persons	266,504	1,066,016
5 or more persons	127,653	670,528
Total	1,857,993	4,131,874

Note that the average family size for families with 5 or more persons is $670,528/127,653 = 5.25$. We use 5.25 as an estimate of the average household size for households of size 5 or more, and divide by 5.25 instead of 5 in all estimates of H_5 .

5.1 Maximum Likelihood Estimation and Poststratification

The estimated household distributions are presented in table 5. The estimates are based on the maximum likelihood (m.l.) estimator (4.1) using the population model with the restricted parametric link function $p_{y,x}$ in combination with the response models RM1(y, z) and RM2(y, z). To illustrate the effect of nonresponse modeling versus post-stratification we also present the standard poststratified estimator (4.4). We recall that this is the maximum likelihood estimator when ignoring the response mechanism. Furthermore, we present the estimated household size distribution based on the imputation-based poststratification (4.3) with the saturated model (4.9). For assessing the sampling variability of the different estimators, the estimated standard errors are also included.

The three models that take the response mechanism into account give higher total number of households. They also give considerable higher numbers of one-person-households. This seems sensible since we expect the one-person households to have the highest nonresponse rate. And thus, these estimates are most influenced by taking the response mechanism into account. We note that the restricted parametric link model (3.1) together with the logistic response model RM2(y, z) gives practically the same poststratified estimates as model (4.9), with also approximately the same standard errors. Because of the freedom of model (4.9), with perfect fit, it seems that model (3.1) & RM2(y, z) works well for estimating the number of households of different sizes. Regarding the uncertainty of the estimates, we see as one might expect that the standard errors typically seem to increase with the number of unknown parameters in the underlying model. Also, the total number of households is rather accurately estimated, not counting possible bias, while it's clearly most difficult to estimate the number of one-person households.

In order to evaluate the extent to which the differences between the estimates are due to sampling error or non-response bias, we consider the estimated standard errors of the differences of the point estimates. Some of these are given in table 6, using mostly the imputation-based post-stratification with the saturated model as a reference. For short, we use the terms Est1 – Est4 for the estimates defined as they appear in table 5:

Est1: M.I. estimator based on population model $p_{y,x}$ and response model RM1

Est2: M.I. estimator based on population model $p_{y,x}$ and response model RM2

Est3: Imputation-based poststratification based on the saturated model (4.9)

Est4: Poststratified estimator without imputation.

Based on tables 5 and 6 we can conclude that Est4 and Est3 have different expected values in estimating H_1 , H_3 , H_5 and H . Regarding the other comparisons, we see that in estimating H_3 there is a significant difference between Est1 and Est2/Est3, and note from earlier discussions in Section 3.3 that RM2 gives a better fit to the data than RM1.

The estimates based on the expansion estimator $\hat{H}_{y,e}$, given by (4.5), in 100's, are 390,500, 496,500, 283,900, 279,900, 148,000 and 1,598,800 with estimated standard errors equal to 33,100, 21,700, 14,600, 11,600, 6,100 and 23,700 for H_1 , ..., H_5 and H , respectively. The standard errors for the differences between these estimates and the Est3-estimates are 52,800, 30,900, 19,100, 10,800, 5,400 and 32,000 for H_1 , ..., H_5 and H respectively. These expansion estimates indicate serious bias due to non-response, especially the estimates for H_1 , H_5 and H ,

with poststratification correcting for some of the bias (probably about 50% for the estimates of H_1 and H). We also note that the standard errors for the poststratified estimator and this simple expansion estimator are about the same. So by reducing the bias with poststratification one reduces the total error as well.

Poststratification corrects for the bias caused by the discrepancy between the family size distributions in the response sample and the population. From table 1 and table 4 we see that these family size distributions are given by (in percentages), for $x = 1, \dots, 5$:

Response sample: 14.6 – 20.7 – 19.1 – 27.0 – 18.6
Population: 19.2 – 19.8 – 19.0 – 25.8 – 16.2.

Since the number of one-person families is much too low in the response sample, so will the expansion estimate of H_1 be. With post strata determined by family size, post-stratification corrects for the family size bias in the response sample, but does implicitly assume that nonrespondents and respondents have the same household size distribution, for a fixed family size. Or, in other words, the respondents are treated as a random subsample of sampled units with the same family size, as mentioned by Little (1993). This is most likely not the case. We recall that the family size variable was not significant when the household variable was included in the response models. Thus it seems reasonable to assume, as in our response models, that response rates will vary with the actual household sizes rather than the registered family sizes. Typically, estimates of the number of one-person households will be biased when the nonrespondents are ignored.

Table 5

Estimated Household Totals for Persons Aged Less than 80 Years in Norway at January 1, 1993, in Units of 100.
In Parentheses, the Estimated Standard Error of the Estimates

Household size, y	Maximum likelihood estimator with nonignorable response mechanism				Imputation-based poststratification		Ignoring the response mechanism	
	Population model $p_{y,x}$ and response model RM1 (y, z)	%	Population model $p_{y,x}$ and response model RM2 (y, z)	%	Saturated population and response model	%	Poststratified estimator	%
1	558,800 (38,900)	32	595,400 (48,000)	34	596,600 (53,500)	34	486,000 (35,800)	29
2	520,200 (20,600)	30	525,800 (27,400)	30	523,600 (29,800)	30	507,800 (20,000)	30
3	278,900 (13,800)	16	249,100 (20,300)	14	250,000 (19,800)	14	286,200 (14,100)	17
4	258,900 (9,800)	15	269,000 (11,600)	15	268,900 (11,500)	15	270,600 (10,100)	16
≥ 5	125,800 (4,700)	7	126,000 (5,100)	7	126,200 (5,000)	7	131,300 (4,700)	8
Total	1,742,600 (25,600)	100	1,765,300 (29,700)	100	1,765,300 (31,900)	100	1,681,900 (23,300)	100

Table 6
Estimated Standard Errors of the Differences of the Point Estimates in Table 5

Household size	Est1–Est2	Est1–Est3	Est2–Est3	Est4–Est3
1	29,700	37,000	16,600	42,400
2	19,300	22,200	8,800	23,100
3	15,400	15,200	5,300	15,500
4	6,700	6,500	1,800	6,600
≥ 5	1,700	1,700	500	1,900
Total	15,300	18,800	8,900	23,300

After having corrected for nonresponse bias by completing the sample with imputed values, the sample itself may be skewed compared to the population. To illustrate the effect of poststratification to correct for this, we shall compare, using the saturated model (4.9), the imputation-based poststratified estimates Est3 with the imputation-based expansion estimates given by (4.6): 583,900, 567,700, 244,300, 259,300, 122,400 and 1,777,600 for H_1 , ..., H_5 and H , respectively. As noted in Section 4.2, see (4.14), these estimates are identical to the modified Horvitz-Thompson estimates. The standard errors for these estimates are practically the same as for Est3. Hence, the alternative poststratified estimation methods based on nonignorable response models have standard errors at least no worse than the modified Horvitz-Thompson estimator. So if one reduces the bias with the alternative methods, one reduces the total error too. The standard errors of the differences between Est3 and this modified Horvitz-Thompson estimator in the estimates of H_1 , ..., H_5 and H are 3,500, 2,200, 1,100, 600, 200 and 2,100 respectively. Clearly these two methods give significantly different estimates for all household size totals. In this comparison, one feature stands out. The expansion estimate of the number of two-persons households, 567,700, is clearly too high, as seen by comparing the family size distributions in the total sample and the population (in percentages), for $x = 1, \dots, 5$:

Population:	19.2 – 19.8 – 19.0 – 25.8 – 16.2
Sample:	18.6 – 23.0 – 17.8 – 24.9 – 15.7.

The sample proportion of persons in two-persons families is much too high, and even though we have corrected for nonresponse bias, the expansion estimator, and then also the modified Horvitz-Thompson estimator cannot correct for a nonrepresentative sample. This will necessarily lead to biased estimates of H_2 . We need poststratification to correct for a skewed sample. One can regard the difference in expected values for these estimators of H_2 as being close to the bias for the modified Horvitz-Thompson estimator, and note that an approximate 95% confidence interval for this difference is (39,800, 48,400).

For robustness considerations we also present the estimates from the cumulative logit model mentioned in Section 3.1 together with RM1 (y, z), which we know fits the

data poorly. They are, in 100's: 591,800, 501,000, 265,200, 267,300, 128,200 and 1,753,500 for H_1 , ..., H_5 and H , respectively. Compared to table 5, this seems to indicate that a reasonable model for response plays a more important role than a good population model. It is also evident that nonresponse modeling makes a difference, as seen when compared to poststratification and simple expansion.

5.2 Comparison with the Currently Used Estimates in CES, the Quality Survey for the 1990 Census and a Projection Study

Since 1993, an alternative, computationally simpler, modified Horvitz-Thompson estimator of type (4.14) has been in use in the production of official statistics from CES, see (Belsby 1995). We recall from Section 2 that the weights are the inverse sampling probabilities of the households, multiplied with the estimated probability of response. The response probabilities are estimated using a logistic model similar to RM2 (y, z) with place of residence and household size as explanatory variables. For the nonrespondents with unknown household size the registered family size is used instead, replacing (3.5). Thus, the weights may be regarded as an approximation to using (3.5). Of course, (3.5) is possible only when a population model is considered, which CES has not done. Table 7 presents estimated household distribution based on this CES-modified Horvitz-Thompson estimator.

The quality survey for the Census 1990, PES 1990, contains 8,280 respondents and uses practically the same household definition as CES. The response rate was 95%. The H_y -estimates uses poststratification with respect to household size in the Census. However, no attempts were made to correct for possible nonresponse bias with respect to actual household size. PES deals with the whole population. Table 7 has the estimates for the 0–79 age group with the same poststratification method as in PES.

Table 7 also presents estimates based on the Household Projections study by Keilman and Brunborg (1995). This study simulates household structure for the period 1990 to 2020. The data sources are 28,384 individuals from the 1990 Population and Housing Census and 1988 Family and Occupation Survey. Keilman and Brunborg project for the whole population in 1992. We adjust their estimates to the 0–79 age group.

Table 7
Estimated Household Size Totals for Persons Less than 80 Years in Norway at January 1, 1993
with CES-modified Horvitz-Thompson, PES 1990 and Projections, in Units of 100

Household size	CES-Modified Horvitz-Thompson	%	PES 1990	%	Projections	%
1	622,900	35	626,000	35	668,300	37
2	518,500	29	494,200	28	549,000	30
3	259,900	15	291,500	16	211,900	12
4	258,500	15	250,000	14	221,500	12
≥ 5	124,600	7	115,300	6	97,500	5
Unknown					78,500	4
Total	1,784,400	1	1,777,000	99	1,826,700	100

Table 8
Estimated Probability of Response Based on the Method Used
in CES Since 1993, in Percentages

Place of residence	Household size				
	1	2	3	4	5 or more
			CES-method		
Rural	44.53	66.24	74.55	73.54	80.07
Urban	36.01	57.90	67.25	66.09	73.80
		Model $p_{y,x}$ in (3.1) combined with RM2(y, z)			
Rural	47.77	60.90	79.05	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46

The estimates in table 7 support our impression that the estimates based on modeling the response mechanism leads to less biased estimates compared with ignoring the response mechanism as in mere poststratification or simple expansion. This is especially true for the one-person households and the total. The current "official estimator", the modified Horvitz-Thompson seems to give estimates of the right magnitude and in fact is closer to the results of PES 1990 than the modelbased estimates. However, this is more by accident. As a *method* it has some problems even in a representative sample. We can study this by estimating the response probabilities. Table 8 presents the results together with the estimates based on RM2(y, z) & (3.1) from table 3.

Compared to the estimated response probabilities based on model RM2(y, z) with (3.1), we see that replacing household size with family size in the nonresponse group is not a satisfactory approximation. Hence, if compared with the modified Horvitz-Thompson estimator in Section 5.1 based on the saturated model (4.9), the latter one would be preferred. For this particular survey, the CES approach overestimates the probability of response for household of size 2, which in a representative sample would lead to underestimating of H_2 . The estimated response probabilities will most likely be biased when we are using family size in place of household size in the nonresponse group when estimating the parameters in the response model. This bias is an additional problem to the previously mentioned one, that the modified Horvitz-Thompson estimates will be

similar to the imputation-based expansion estimates and cannot correct for nonrepresentative samples (as has been a problem in CES since 1993). In the 1992 CES, however, the sample is skewed with a too high proportion of families of size 2, and the H_2 -estimate will be of the right magnitude, by accident.

6. Conclusions

We have investigated modeling and methodological issues for estimating the total number of households of different sizes in Norway, based on the Norwegian Consumer Expenditure Survey (CES). The main issue is how to correct for bias due to nonignorable nonresponse. The existing estimation method in CES is a modified Horvitz-Thompson estimator that includes a correction for nonresponse by estimating response probabilities. We have considered basically two modelbased approaches, a maximum-likelihood estimator and imputation-based post-stratification after registered family size. With a population model that corresponds to a group model after family size only, these two estimators are identical. This family group model for household size and a logistic link for the response probability using household size as a categorical variable seem to work well for our estimation problem.

In analyzing the 1992 CES, we find serious bias due to nonresponse, especially the estimates for H_1 and H_2 with pure poststratification (without imputation) correcting for

some of the bias (probably about 50% for the estimates of H_1 and H). Poststratification does not, however, take into account possible nonresponse bias dependent on household size. Our response models assume that the response rates will vary with the actual household sizes rather than the registered the family sizes, and it is quite evident that such nonresponse modeling makes a difference, leading to less biased estimates than mere poststratification or simple expansion, especially of H_1 and H .

The modified Horvitz-Thompson estimates used in the official statistics from CES correspond to imputation-based expansion estimates. Hence, they cannot correct for nonrepresentative samples. The study in this paper shows that, in addition to a nonignorable response model it is also necessary to poststratify according to family size, *i.e.*, using a population model given family size. Hence poststratification, response modeling and imputation are key ingredients for a satisfactory approach.

In any estimation problem of totals in survey sampling, one must be aware of the fact that a Horvitz-Thompson estimator cannot correct for skewed samples, even when modified with good response estimates. Poststratification should always be considered as well as imputation based on a response model, nonignorable when needed.

Appendix A1

The data for rural and urban areas separately are given in table A1.

Appendix A2

Theorem. Assume model (3.1) for Y . *i.e.*, $P(Y = y | x, z) = p_{y,x}$ is independent of z , but otherwise the $p_{y,x}$'s are completely unknown with the only restriction being that $\sum_y p_{y,x} = 1$, for all values of x , for all k . The response mechanism is arbitrarily parametrized, *i.e.*, no

assumption is made about $P(R=1 | Y = y, x, z)$. Then the maximum likelihood estimates for $p_{y,x}$ are given by

$$\hat{p}_{y,x} = \frac{n_{xy} + n_{xy}^*}{m_x + m_{xu}}.$$

Proof. Let $q_{y,x,z} = P(R=1 | Y = y, x, z)$. The log likelihood is given by

$$\begin{aligned} \ell &= \sum_x \sum_y n_{xy} p_{y,x} + \sum_{z=0}^1 \sum_x \sum_y n_{xy}(z) q_{y,x,z} \\ &\quad + \sum_{z=0}^1 \sum_x m_{xu}(z) \log P(R=0 | x, z) \\ &= \sum_x \sum_y n_{xy} p_{y,x} + \sum_{z=0}^1 \sum_x \sum_y n_{xy}(z) q_{y,x,z} \\ &\quad + \sum_{z=0}^1 \sum_x m_{xu}(z) \log(1 - \sum_{y=1}^5 p_{y,x} q_{y,x,z}). \end{aligned}$$

We use the Lagrange method and maximize $G = \ell + \sum_{x=1}^5 \lambda_x (\sum_{y=1}^5 p_{y,x} - 1)$.

Let the solutions be $\hat{p}_{y,x}(\lambda_x)$, and determine the λ_x 's such that $\sum_y \hat{p}_{y,x}(\lambda_x) = 1$, for all x . No matter how the $q_{y,x,z}$'s are parametrized, the mle $\hat{p}_{y,x}$ must satisfy, by solving the equations $\partial G / \partial p_{y,x} = 0$,

$$\frac{n_{xy}}{\hat{p}_{y,x}} - \sum_{z=0}^1 m_{xu}(z) \frac{\hat{q}_{y,x,z}}{\hat{P}(R=0 | x, z)} + \lambda_x = 0 \quad (\text{A1})$$

which is equivalent to:

$$\begin{aligned} n_{xy} &= \hat{p}_{y,x} \sum_{z=0}^1 \frac{m_{xu}(z)}{\hat{P}(R=0 | x, z)} \\ &\quad - \sum_{z=0}^1 m_{xu}(z) \frac{\hat{P}(R=0, Y=y | x, z)}{\hat{P}(R=0 | x, z)} - \hat{p}_{y,x} \lambda_x. \end{aligned}$$

Table A1

Family and Household Sizes for the 1992 Norwegian Consumer Expenditure Survey, Split into Rural and Urban Areas. The Upper Entry is for the Urban Group

Family size	Household size					Total response	Non-response	Total	Response rate
	1	2	3	4	≥ 5				
1 urban	28	24	7	2	0	61	78	139	0.439
rural	55	24	13	7	2	101	75	176	0.574
2 urban	6	70	12	3	0	91	84	175	0.520
rural	3	107	25	1	3	139	76	215	0.647
3 urban	4	8	57	11	3	83	40	123	0.675
rural	6	17	74	29	3	129	51	180	0.717
4 urban	0	3	15	80	5	103	43	146	0.705
rural	2	10	22	151	12	197	80	277	0.711
≥ 5 urban	0	1	0	6	66	73	28	101	0.723
rural	1	3	4	11	115	134	32	166	0.807
Total urban	38	106	91	102	74	411	273	684	0.601
Total rural	67	161	138	199	135	700	314	1014	0.690

We determine λ_x by summing over y :

$$m_x = \sum_{z=0}^1 \frac{m_{xu}(z)}{\hat{P}(R=0|x,z)} - \sum_{z=0}^1 m_{xu}(z) \frac{\hat{P}(R=0|x,z)}{\hat{P}(R=0|x,z)} - \lambda_x,$$

hence

$$\lambda_x = \sum_{z=0}^1 \frac{m_{xu}(z)}{\hat{P}(R=0|x,z)} - (m_x + m_{xu}).$$

It follows from (A1) that $\hat{p}_{y,x}$ satisfies the following relation:

$$\hat{p}_{y,x} = \frac{n_{xy}}{\left(m_x + m_{xu} - \sum_{z=0}^1 m_{xu}(z) \frac{\hat{P}(R=0|Y=y, x, z)}{\hat{P}(R=0|x,z)}\right)}. \tag{A2}$$

The imputed values are given by , from (4.2),

$$n_{xy}^*(z) = m_{xu}(z) \hat{p}_{y,x} \frac{\hat{P}(R=0|Y=y, x, z)}{\hat{P}(R=0|x,z)}$$

and, from (A2),

$$\begin{aligned} \hat{p}_{y,x} &= n_{xy} \bigg/ \left(m_x + m_{xu} - \sum_{z=0}^1 \frac{n_{xy}^*(z)}{\hat{p}_{y,x}} \right) \\ &= n_{xy} \bigg/ \left(m_x + m_{xu} - \frac{n_{xy}^*}{\hat{p}_{y,x}} \right) \end{aligned}$$

or equivalently,

$$\hat{p}_{y,x} (m_x + m_{xu}) - n_{xy}^* = n_{xy},$$

$$\text{i.e., } \hat{p}_{y,x} = \frac{n_{xy} + n_{xy}^*}{m_x + m_{xu}}. \quad \text{Q.E.D}$$

Appendix A3

Table A2

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (3.1) and RM1 (y, z)

Family size		Household size					Total
		1	2	3	4	≥ 5	
1	urban	77.8	44.1	12.9	3.9	0.3	139
	rural	103.6	43.1	18.4	8.7	2.3	176
2	urban	10.8	137.9	22.1	3.8	0.4	175
	rural	7.5	168.6	33.9	1.7	3.3	215
3	urban	7.5	14.3	81.3	16.4	3.6	123
	rural	10.7	25.3	104.8	35.6	3.7	180
4	urban	0.8	6.4	21.9	110.3	6.6	146
	rural	3.5	16.7	35.1	206.9	14.8	277
≥ 5	urban	0.5	2.4	1.0	9.0	88.2	101
	rural	1.6	4.7	5.2	14.4	140.1	166
Total /urban		97.4	205.1	139.2	143.4	99.1	684
rural		126.9	258.4	197.4	267.3	164.2	1,014

Table A3

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (3.1) and RM2 (y, z)

Family size		Household size					Total
		1	2	3	4	≥ 5	
1	urban	81.6	42.7	10.4	4.0	0.3	139
	rural	107.5	41.5	15.9	8.8	2.3	176
2	urban	11.9	140.4	18.3	3.9	0.5	175
	rural	8.6	170.9	30.3	1.8	3.4	215
3	urban	9.4	16.1	75.2	18.6	3.7	123
	rural	13.4	27.7	96.5	38.5	3.9	180
4	urban	0.8	6.2	18.9	113.5	6.6	146
	rural	3.7	16.2	29.2	213.1	14.8	277
≥ 5	urban	0.5	2.3	0.6	9.3	88.3	101
	rural	1.7	4.6	4.6	14.9	140.2	166
Total /urban		104.2	207.7	123.4	149.3	99.4	684
rural		134.9	260.9	176.5	277.1	164.6	1,014

Appendix A4

Table A4

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban.
The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group.
Based on Model (4.9), i.e., Imputations Determined by (4.7) and (4.8)

Family size	Household size					Total
	1	2	3	4	≥ 5	
1 urban	79.6	47.2	9.4	2.8	0.0	139
rural	108.3	38.5	16.9	9.9	2.4	176
2 urban	17.1	137.7	16.0	4.2	0.0	175
rural	5.9	171.6	32.5	1.4	3.6	215
3 urban	11.4	15.7	76.2	15.6	4.1	123
rural	11.8	27.3	96.2	41.1	3.6	180
4 urban	0.0	5.9	20.0	113.2	6.9	146
rural	3.9	16.0	28.6	214.0	14.5	277
≥ 5 urban	0.0	2.0	0.0	8.5	90.5	101
rural	2.0	4.8	5.2	15.6	138.4	166
Total /urban	108.1	208.5	121.6	144.3	101.5	684
rural	131.9	258.2	179.4	282.0	162.5	1,014

Table A5

The Total Numbers of Family and Household Sizes for Imputed Complete Sample. Based on Model (4.9)

Family size	Household size					Total
	1	2	3	4	≥ 5	
1	187.9	85.7	26.3	12.7	2.4	315
2	23.0	309.2	48.6	5.7	3.6	390
3	23.2	43.0	172.4	56.7	7.7	303
4	3.9	21.9	48.7	327.2	21.3	423
≥ 5	2.0	6.8	5.2	24.1	229.0	267
Total	240.0	466.6	301.1	426.3	264.0	1,698

References

- Baker, S.G., and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.
- Belsby, L. (1995). Forbruksundersøkelsen. Vektmetoder, frafallskorrigerer og intervjuer-effekt. (The consumer survey. Weight methods, nonresponse correction and interviewer effect), Notater 95/18 Statistics Norway.
- Bjørnstad, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.
- Bjørnstad, J.F., and Skjold, F. (1992). Interval estimation in the presence of nonresponse. *The American Statistical Association 1992 Proceedings of the Section on Survey Research Methods*, 233-238.
- Bjørnstad, J.F., and Walsøe, H.K. (1991). Predictive likelihood in nonresponse problems. *The American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*, 152-156.
- Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.
- Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (with discussion). *Journal of the Royal Statistical Society B*, 60, 57-70.
- Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- Hall, B.H., Cummins, C. and Schnake, R. (1991). *TSP Reference Manual, Version 4.2A*, Palo Alto California: TSP International.
- Holt, D., and Smith, T.M.F. (1979). Post-stratification, *Journal of the Royal Statistical Society A*, 142, 33-46.
- Keilman, N., and Brunborg, H. (1995). *Household Projections for Norway, 1990-2020, Part 1: Macrosimulation*, Reports 95/21, Statistics Norway.
- Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- Little, R.J.A., and Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

- McCullagh, P., and Nelder, J.A. (1991). *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.
- Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Statistics Norway (1990). *Survey of Consumer Expenditure 1986-88*. Official Statistics of Norway NOS B919.
- Statistics Norway (1996). *Survey of Consumer Expenditure 1992-1994*. Official Statistics of Norway NOS C317.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Bayesian Analysis of Nonignorable Missing Categorical Data: An Application to Bone Mineral Density and Family Income

Balagobin Nandram, Lawrence H. Cox and Jai Won Choi¹

Abstract

We consider a problem in which an analysis is needed for categorical data from a single two-way table with partial classification (*i.e.*, both item and unit nonresponses). We assume that this is the only information available. A Bayesian methodology permits modeling different patterns of missingness under ignorability and nonignorability assumptions. We construct a nonignorable nonresponse model which is obtained from the ignorable nonresponse model via a model expansion using a data-dependent prior; the nonignorable nonresponse model robustifies the ignorable nonresponse model. A multinomial-Dirichlet model, adjusted for the nonresponse, is used to estimate the cell probabilities, and a Bayes factor is used to test for association. We illustrate our methodology using data on bone mineral density and family income. A sensitivity analysis is used to assess the effects of the data-dependent prior. The ignorable and nonignorable nonresponse models are compared using a simulation study, and there are subtle differences between these models.

Key Words: Bayes factor; Chi-squared statistic; Importance function; Markov chain Monte Carlo; Multinomial-Dirichlet model; Robust; Two-way categorical table.

1. Introduction

It is a common practice to use two-way categorical tables to present survey data. For many surveys there are missing data, and this gives rise to partial classification of the sampled individuals. Thus, for the two-way table there are both item nonresponse (one of the two categories is missing) and unit nonresponse (both categories are missing); see Little and Rubin (2002, section 1.3) for definitions of the three missing data mechanisms (MCAR, MAR, MNAR). Thus, there are four tables (one table with the complete cases, and three possible supplemental tables: one table with row classification only, one table with column classification only, and one table with neither row nor column classification). One may not know how the data are missing. Thus, we use a model in which the likelihood function accounts for differences between the observed data and missing data (*i.e.*, nonignorable missing data); see Rubin (1976) and Little and Rubin (2002) for the relation between ignorability/nonignorability and these three missing data mechanisms. Because there are well-known advantages of the Bayesian method over the non-Bayesian method for these problems, we propose a Bayesian analysis of a general $r \times c$ categorical table, consisting of a table with complete cases and three supplemental tables. Specifically, we develop a Bayesian method to estimate the cell probabilities and test for association between the two categorical variables.

We assume that the only information available to the data analysts is the complete cases and the three supplemental tables. Specifically, we assume that there is no information (either from covariates or prior information) about nonignorability. In our Bayesian approach, the survey design features have been suppressed (*i.e.*, there are no survey weights and there are no clustering or stratification). Sometimes survey data are presented to the public with certain features of the data suppressed for reasons of convenience and confidentiality. We recognize that both the ignorable and the nonignorable nonresponse models may be incorrect when they do not take account of these features. However, the parameters in the ignorable nonresponse model are identifiable and estimable, and one can take advantage of this fact to construct a nonignorable nonresponse model which is related to the ignorable nonresponse model. Also, in the ignorable nonresponse model we assume that there is a MAR mechanism that drives the nonresponse, and there may be information in the incomplete cases (*i.e.*, the two tables with observed row and column margins). Without any information about the degree of nonignorability, it is sensible to generalize the ignorable nonresponse model. This is how we attempt to accomplish our objectives in this paper.

This paper has five sections. In section 1 we have further discussion of the problem, and we review related methodology. In section 2, we describe a 3×3 table of bone mineral density (BMD) and family income (FI) from the third National Health and Nutrition Examination Survey

1. Balagobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute road, Worcester MA 01609, E-mail: balnan@wpi.edu; Lawrence H. Cox and Jai Won Choi, Office of Research and Methodology, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. E-mail: lgc9@cdc.gov, jwc7@cdc.gov.

(NHANES III). This is used mainly for illustration. In section 3, we describe the methodology to obtain estimates of the cell probabilities, and we use the Bayes factor to test for association of the two attributes. We accomplish these objectives by first constructing an ignorable nonresponse model, and we show how to expand an ignorable nonresponse model into a nonignorable nonresponse model. In section 4, we analyze the NHANES III data to demonstrate our methods. Also, a simulation study gives further comparison of the ignorable and the nonignorable nonresponse models, and a sensitivity analysis shows that inference is not too sensitive to the choice of an important prior distribution. Finally, section 5 has concluding remarks.

1.1 Discussion of the Problem

We do not know whether an ignorable nonresponse model or a nonignorable nonresponse model is appropriate, but it is worthwhile noting that Cohen and Duffy (2002) point out that "Health surveys are a good example, where it seems plausible that propensity to respond may be related to health." Thus, nonignorable nonresponse models are important candidates for the analysis of data from health surveys. For a general $r \times c$ categorical table (two categorical variables, one with r categories and the other with c categories) with nonresponse, our objectives are to show how to (a) make inference about the cell probabilities, and (b) test for no association between the two categories using the Bayes factor. While (a) comes directly from the modeling, (b) needs one extra step.

Let I_i be the cell indicator for the i^{th} individual in a $r \times c$ table for $i=1, \dots, n$ individuals. Then, it is well known that if the I_i are *independent and identically* distributed, the Pearson's chi-squared statistic has $\chi^2_{(r-1)(c-1)}$. Otherwise the Pearson's chi-squared statistic does not have a $\chi^2_{(r-1)(c-1)}$, and this is true when there are missing data and the respondents and nonrespondents differ. When this is the case, adjustments must be made to the Pearson's chi-squared statistic. Within the non-Bayesian framework Chen and Fienberg (1974) and Wang (2001) have corrections for incomplete two-way tables. Although not directly relevant here, it is pertinent to mention that similar adjustments have been made for cluster sampling and stratified random sampling (Rao and Scott 1981, 1984). The works of Chen and Fienberg (1974) and Wang (2001) can essentially handle item nonresponse only; unit nonresponse is excluded because the modeling is motivated by the ignorable nonresponse models (e.g., see discussion in Kalton and Kasprzyk 1986).

The Bayesian method permits us to use a procedure that does not rely on asymptotic theory, incorporate nonignorable missingness into the modeling and obtain an alternative to Pearson's chi-squared statistic for testing for

no association; see Little (2003) for a discussion of the well-known advantages of the Bayesian approach in survey sampling. Our alternative to the Pearson chi-squared statistic is based on the Bayes factor (Kass and Raftery 1995). This is a statistic that compares a model with association and one with no association via the ratio of their marginal likelihoods under the ignorable and the nonignorable nonresponse models separately.

Little and Rubin (2002, chapter 15) discuss the nonignorable nonresponse problem. For example, Rubin, Stern and Vehovar (1995) (also discussed in Little and Rubin 2002, page 345) provide an interesting analysis of the November/December 1990 Slovenian Public Opinion survey in which there were data on 2,074 prospective voters in their plebiscite with three dichotomous variables; there is 12% nonresponse. They fit both ignorable and nonignorable nonresponse models (loglinear with all interactions) to the data, and they were satisfied with the ignorable nonresponse model. However, they stated "Of course, this does not mean that MAR should be automatically applied in all cases. Analyses assuming MAR are not likely to be adequate if a survey has large amounts of nonresponse, if covariate information is limited, or for cases where the missing-data mechanism is clearly nonignorable (e.g., censored data)."

1.2 Related Methodology

Our methodology is different from Rubin, Stern and Vehovar (1995). We start with Nandram and Choi (2002 a, b) in which a parameter γ centers (can be viewed as an index) the nonignorable nonresponse model on the ignorable nonresponse model. When $\gamma=1$, the nonignorable nonresponse model is the ignorable nonresponse model, and thus, the nonignorable nonresponse model "degenerates" into the ignorable nonresponse model when $\gamma=1$; see also Forster and Smith (1998). This is useful because the nonignorable nonresponse model contains the ignorable nonresponse model as a special case; thereby expressing uncertainty about ignorability. Draper (1995) called this a *continuous model expansion*, and he has recommended the use of a continuous model expansion over a discrete model expansion (i.e., finite mixtures) whenever it is possible. We simply call the continuous model expansion an *expansion model*. Nandram and Choi (2002 a, b) obtain the centering by taking $\gamma|v \sim \text{Gamma}(v, v)$ in which $E(\gamma|v)=1$, $\text{var}(\gamma|v)=1/v$.

Nandram and Choi (2002 a) analyze binary data on household crimes in the National Crime Survey, and Nandram and Choi (2002 b) analyze binary data on doctor visits in the National Health Interview Survey. While Nandram and Choi (2002 a) has more comparisons, Nandram and Choi (2002 b) has more sensitivity analyses. Nandram, Han and Choi (2002) describe two hierarchical

Bayesian models, an ignorable and a nonignorable non-response model, for the analysis of count data from several areas, the counts in each area being described by a multinomial distribution. In all these works the issue of association is not relevant because there is a single categorical variable.

The approach in Nandram and Choi (2002 a, b) is attractive, but it does not apply immediately to the current application on $r \times c$ categorical table. Specifically, only one centering parameter was needed in Nandram and Choi (2002 a, b). To extend the method of Nandram and Choi (2002 a, b), one needs rc centering parameters. Each of these parameters has to have a distribution centered at one to allow degeneration to the ignorable nonresponse model. There are also inequality constraints that must be included in the nonignorable nonresponse model. Thus, while this idea is attractive, the methodology needed to apply the work of Nandram and Choi (2002 a, b) is much beyond the scope of our current paper.

Nandram, Liu, Choi and Cox (2005) extend the work of Nandram, Han and Choi (2002) in two important directions to (a) consider several two-way categorical tables instead of one-way tables and (b) develop a method to study the association between the two categorical variables. Nandram, Liu, Choi and Cox (2005) analyze data on the relation between bone mineral density (BMD) and age from thirty-five counties in the third National Health and Nutrition Examination Survey. In each county the data are categorized into two levels of age and three levels of BMD (*i.e.*, there are thirty-five 2×3 categorical tables). Note that the age of everyone is observed, but the BMD values for a large number of individuals are not observed. Thus, for each county there is a single table with complete cases, and one table with row totals (*i.e.*, the ages of these individuals are known, but their BMD values are missing). Here, our objective is to extend the work of Nandram, Liu, Choi and Cox (2005) to a *general* $r \times c$ categorical table. This is an important advance because now there are three supplemental tables (one table with row classification only, one table with column classification only, and one table with neither row nor column classification) instead of just one with row totals as in Nandram, Liu, Choi and Cox (2005).

2. Data on Bone Mineral Density and Family Income

We briefly describe the 3×3 categorical table of bone mineral density (BMD) and family income (FI). FI is a discrete variable, and there are three levels: low, medium and high. While BMD is a continuous variable, the World Health Organization has classified BMD into three levels: normal, osteopenia and osteoporosis; see Looker, Orwoll,

Johnston, Lindsay, Wahner, Dunn, Calvo and Harris (1997, 1998). BMD is used to diagnose osteoporosis, a disease of elderly females, and in NHANES III it is measured for individuals at least twenty years old (*i.e.*, we use the data on white females only with chronic conditions older than twenty years).

Among those participated in the examination stage, about 62% of the individuals have both FI and BMD observed, 8% with only BMD observed, 29% with only income observed, 1% with neither income nor BMD. The dataset, used in our study, is presented in Table 1 as a 3×3 categorical table of BMD and FI. Our problem is to estimate the proportion of individuals at each BMD-FI level and to test for association between BMD and FI. In NHANES III the response rate increases up to age twenty years, and stabilizes after that age; race, sex and the sampling weights play a minor role (see Nandram and Choi 2005). Thus, for this application we assume that the only data available are the four tables of BMD and FI, and we develop a methodology for this situation.

Table 1

Classification of Bone Mineral Density (BMD) and Family Income (FI) for 2,998 White Females, at Least 20 years Old (20+)

BMD	FI				Sum
	0	1	2	Missing	
0	621	290	284	135	1,330
1	260	131	117	69	577
2	93	30	18	27	168
Missing	456	156	266	45	923
Sum	1,430	607	685	276	2,998

Note: BMD: 0(> 0.82g/cm²; normal), 1(> 0.64, ≤ 0.82g/cm²; osteopenia), 2(≤ 0.64g/cm²; osteoporosis); FI: 0(< \$20,000), 1(≥ \$20,000, < \$45,000), 2(≥ \$45,000); BMD is only measured for age 20+.

It is difficult to assess an association between BMD and FI when there are many individuals not completely classified (*i.e.*, missing data). As discussed in the literature, not necessarily on NHANES III, there are several potentially important confounding variables such as age, smoking, dietary calcium intake, estrogen replacement therapy, physical activity, educational attainment, health status and alcohol consumption (see Ganry, Baudoin and Fardellone 2000). Farahmand, Persson, Michaelsson, Baron, Parker and Ljunghall (2000) stated that for postmenopausal women, aged 50–81 years, from six counties in Sweden, higher household income is associated with decreased hip fracture risk. Using complete data from NHANES III, Lauderdale and Rathouz (2003) studied the regression of bone mineral content on economic indicators (*e.g.*, education and poverty income ratio). An adjustment was made for other factors such as age, height and weight. They conclude that “Bone density does not reflect economic conditions as

strongly or consistently as physical stature." Unfortunately, these works do not address the nonignorability of the missing data; missing data are not discussed. Also, the response rate to income items is usually low.

We have looked at the data for the complete cases more closely. We fit a multinomial-Dirichlet model with association and one with no association. The model with association is $\mathbf{n} | \mathbf{p} \sim \text{Multinomial}(\mathbf{n}, \mathbf{p})$ and $\mathbf{p} \sim \text{Dirichlet}(1, \dots, 1)$. Note that by no association we mean that $p_{jk} = p_j^{(1)} p_k^{(2)}$, $j = 1, \dots, r$, $k = 1, \dots, c$, where $\sum_{j=1}^r p_j^{(1)} = 1$ and $\sum_{k=1}^c p_k^{(2)} = 1$. Thus, for the model with no association, $\mathbf{n} | \mathbf{p} \sim \text{Multinomial}(\mathbf{n}, \mathbf{p})$, $\mathbf{p}^{(1)} \sim \text{Dirichlet}(1, \dots, 1)$, and independently $\mathbf{p}^{(2)} \sim \text{Dirichlet}(1, \dots, 1)$, where $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ have r and c components respectively. It is easy to show that the marginal likelihood with association (as) is $p_{\text{as}}(\mathbf{n}) = (rc-1)!n!/(n+rc-1)!$ and with no association (nas) is

$$p_{\text{nas}}(\mathbf{n}) = p_{\text{as}}(\mathbf{n}) \frac{(r-1)!(c-1)!}{(rc-1)!} \frac{(n+rc-1)!}{(n+r-1)!(n+c-1)!} \frac{\prod_{j=1}^r n_j! \prod_{k=1}^c n_k!}{\prod_{j=1}^r \prod_{k=1}^c n_{jk}!}.$$

Consider our data in Table 1 again. Under independence (i.e., no association) the observed chi-squared statistic is 12.7 on 4 degrees of freedom with a p -value of 0.013 and the hypothesis of no association is rejected. On the logarithmic scale, the marginal likelihoods are $p_{\text{nas}}(\mathbf{n}) = -46.2$ and $p_{\text{as}}(\mathbf{n}) = -49.6$ resulting in a log Bayes factor of 3.40 for evidence of no association relative to association. Therefore, while the chi-squared test provides strong evidence against no association, the log Bayes factor provides strong evidence for no association. Thus, there is a contradictory evidence for no association. See Mirkin (2001) for a review of interpretations of the chi-squared statistic as a measure of association or independence.

How sensitive is the Bayes factor to the choice of the prior distributions? First, note that the prior density that any reasonable person might use in this problem is the Dirichlet distribution. For the model with association we have selected the prior distributions to be $\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\gamma})$, and for the model with no association $\mathbf{p}^{(1)} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ and independently $\mathbf{p}^{(2)} \sim \text{Dirichlet}(\boldsymbol{\beta})$. Let $n_j^{(1)} = \sum_{k=1}^c n_{jk}$, $j = 1, \dots, r$ and $n_k^{(2)} = \sum_{j=1}^r n_{jk}$, $k = 1, \dots, c$. Then, it is easy to show that the Bayes factor for a test of association versus no association is

$$\text{BF} = \frac{D_{rc}(\mathbf{n} + \boldsymbol{\gamma}) / D_r(\mathbf{n}^{(1)} + \boldsymbol{\alpha}) D_c(\mathbf{n}^{(2)} + \boldsymbol{\beta})}{D_{rc}(\boldsymbol{\gamma}) / D_r(\boldsymbol{\alpha}) D_c(\boldsymbol{\beta})},$$

where $D_r(\cdot)$ refers to the Dirichlet function with r components, etc.; see section 3.1 for notations. Then, we choose

each of the components of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to be κ (e.g., in $p_{\text{as}}(\mathbf{n})$ and $p_{\text{nas}}(\mathbf{n})$, $\kappa = 1$). Sensitivity to the choice of prior distributions can be studied in terms of κ . Here $\kappa = 1$ corresponds to the prior distributions that are usually used in the multinomial-Dirichlet model, and $\kappa = 0.50$, Jeffreys' prior. Thus, we have chosen $\kappa = 0.25, 0.5, 1.0, 1.5, 2, 3$, and the corresponding Bayes factors (log scale) are 4.7, 3.6, 3.4, 3.9, 4.7, 6.6. Thus, while the Bayes factor is sensitive to the choice of the prior distributions, it is not too sensitive. Of course, if there is informative prior information, in which κ is substantially large, it is a different issue.

The Pearson chi-squared statistic is dominated by cells (3, 1) and (3, 3) with squares of the Pearson residuals being 4.61 and 6.15 respectively (the observed chi-squared statistic is 12.7). It is interesting that the Bayes factor tends to smooth this effect out. We have collapsed the two categories, osteopenia and osteoporosis, into a single category. For this 2×3 categorical table, the chi-squared test statistic is 1.7 on 2 degrees of freedom with a p -value of 0.42. The marginal likelihoods are $p_{\text{nas}}(\mathbf{n}) = -28.2$ and $p_{\text{as}}(\mathbf{n}) = -32.0$ resulting in a log Bayes factor of -3.81 . Therefore, both tests suggest no association for this 2×3 table. Thus, based on these data it is hard to believe that there is an association between BMD and FI. The question that now arises is "Can this conclusion change if we take into account the incomplete data?"

3. Methodology and Nonresponse Models

First, we describe the notation. Second, we describe the ignorable nonresponse model. Third, we construct a non-ignorable nonresponse model by expanding the ignorable nonresponse model. Fourth, we discuss the Bayes factor. Finally, we describe how to specify an important prior distribution.

3.1 Notation

For a $r \times c$ categorical table, let $I_{jkl} = 1$ if l^{th} individual falls in the j^{th} row and k^{th} column and 0 otherwise. Also, let $J_{sl} = 1$ if the l^{th} individual falls in table s ($s = 1$: complete cases; $s = 2$: table with row totals; $s = 3$: table with column totals; $s = 4$: table with individuals unclassified), and $J_{sl} = 0$ otherwise, $s = 1, 2, 3, 4$ with $\sum_{s=1}^4 J_{sl} = 1$. The vector $\mathbf{J}_l = (J_{1l}, J_{2l}, J_{3l}, J_{4l})'$ has its components corresponding to the four tables.

Let p_{jk} be the probability that an individual belongs to cell (j, k) of the $r \times c$ table, and let π_{sjk} be the probability that an individual belongs to the s^{th} table, given that cell status (j, k) . For the ignorable nonresponse model $\pi_{sjk} = \pi_s$, but for a nonignorable nonresponse model π_{sjk} depends on at least one of j and k as well. We will also let

\mathbf{p} be the vector p_{jk} , $j=1, \dots, r$, $k=1, \dots, c$, and $\boldsymbol{\pi}_{jk}$ be a vector with components $\{\pi_{sjk}, s=1, \dots, 4\}$, $j=1, \dots, r$, $k=1, \dots, c$.

Then, we take

$$\mathbf{I}_\ell | \mathbf{p} \stackrel{\text{iid}}{\sim} \text{Multinomial}\{1, \mathbf{p}\}, \quad (1)$$

where $\sum_{j=1}^r \sum_{k=1}^c p_{jk} = 1$, $p_{jk} \geq 0$, $j=1, \dots, r$, $k=1, \dots, c$. For the parameters \mathbf{p} we take

$$\mathbf{p} \sim \text{Dirichlet}(1, \dots, 1), \quad p_{jk} \geq 0, \quad \sum_{j=1}^r \sum_{k=1}^c p_{jk} = 1. \quad (2)$$

Henceforth, we will use the notation that a k -dimensional vector, $\mathbf{x} \sim \text{Dirichlet}(c\mathbf{t})$ to mean $f(\mathbf{x}) = (\prod_j x_j^{c_j t_j}) / D_k(c\mathbf{t})$, $x_j \geq 0$, $\sum_{j=1}^k x_j = 1$, where $D_k(c\mathbf{t}) = (\prod_{j=1}^k \Gamma(c_j t_j)) / \Gamma(t)$ is the Dirichlet function with $c_j > 0$, $\sum_{j=1}^k c_j = 1$.

Assumptions (1) and (2) are the same for both the ignorable and nonignorable nonresponse models, and they are standard when there are no missing data.

Let the cell counts be $y_{sjk} = \sum_{\ell=1}^n I_{jkt} J_{st}$, $s=1, 2, 3, 4$ for the four cases. Here y_{1jk} are observed and y_{sjk} , $s=2, 3, 4$ are missing (i.e., latent variables). For y_{1jk} we know that $\sum_{j=1}^r \sum_{k=1}^c y_{1jk} = n_0$, the number of individuals with complete data; for y_{2jk} we know that $\sum_{k=1}^c y_{2jk} = u_j$, where the row margins u_j , $j=1, \dots, r$ are observed; for y_{3jk} we know that $\sum_{j=1}^r y_{3jk} = v_k$, where the column margins v_k , $k=1, \dots, c$ are observed; and for y_{4jk} we know that $\sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w$. Throughout we assume that all inference is conditional on $n_0, \mathbf{u}, \mathbf{v}$ and w , and we will suppress this notation whenever it is understood. Whenever it is convenient, we will use notations such as $\sum_{s,j,k} y_{sjk} \equiv \sum_{s=1}^4 \sum_{j=1}^r \sum_{k=1}^c y_{sjk}$, $\prod_{s,j,k} \pi_{sjk} \equiv \prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}$ and $\mathbf{y}_{(1)} = (y_2, y_3, y_4)$, $\mathbf{y}_{(2)} = (y_1, y_3, y_4)$ etc., where $y_s = (y_{sjk}, j=1, \dots, r, k=1, \dots, c)$, $s=1, 2, 3, 4$. Also, $\sum_{s,j,k}^{4,r,c} y_{sjk} = n$. We will also use $y_{s-} = \sum_{j,k} y_{sjk}$, $y_{jk} = \sum_s y_{sjk}$ and $\mathbf{y} = (y_1, y_2, y_3, y_4)$.

3.2 Ignorable Nonresponse Model

For the ignorable nonresponse model we take

$$\mathbf{J}_\ell | \boldsymbol{\pi} \stackrel{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{\pi}\}. \quad (3)$$

That is, there is no dependence on the cell status of an individual.

Then, the augmented likelihood function for $\mathbf{p}, \boldsymbol{\pi}, \mathbf{y}_{(1)} | \mathbf{y}_1, n_0, \mathbf{u}, \mathbf{v}, w$ is

$$g(\mathbf{p}, \boldsymbol{\pi}, \mathbf{y}_{(1)} | \mathbf{y}_1, n_0, \mathbf{u}, \mathbf{v}, w) \propto \left[\prod_{s=1}^4 \pi_s^{y_{s-}} \right] \left[\prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!} \right], \quad (4)$$

subject to $\sum_{j=1}^r \sum_{k=1}^c y_{1jk} = n_0$, $\sum_{k=1}^c y_{2jk} = u_j$, $j=1, \dots, r$, $\sum_{j=1}^r y_{3jk} = v_k$, $k=1, \dots, c$, and $\sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w$. There

are three interesting features in (4). First, under ignorability the likelihood function separates into two pieces, one that contains the π_s only and the other the p_{jk} , and inference about these two parameters are unrelated. Second, inference about π_s is based only on the observed y_{s-} (i.e., the sufficient statistics for π_1, π_2, π_3 and π_4 are essentially the proportions of cases in the first, second, third and fourth tables respectively). Third, under the ignorable nonresponse model, the u_j and the v_k contain information about the p_{jk} ; w does not contain any information about the p_{jk} . This is easy to show; letting T denote the set $\{(y_2, y_3, y_4) : \sum_{k=1}^c y_{2jk} = u_j, j=1, \dots, r, \sum_{j=1}^r y_{3jk} = v_k, k=1, \dots, c, \sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w\}$, by (4)

$$\sum_{(y_2, y_3, y_4) \in T} \prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!} = w! \prod_{j=1}^r \frac{u_j!}{\left\{ \sum_{k=1}^c p_{jk} \right\}^{u_j}} \prod_{k=1}^c \frac{v_k!}{\left\{ \sum_{j=1}^r p_{jk} \right\}^{v_k}} \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{1jk}}}{y_{1jk}!}.$$

Finally, for the parameters $\boldsymbol{\pi}$ we take

$$\boldsymbol{\pi} \sim \text{Dirichlet}(1, \dots, 1), \quad \pi_s \geq 0, \quad \sum_{s=1}^4 \pi_s = 1. \quad (5)$$

Note that this is a uniform probability density in four-dimensional space, and there are no hyperparameters in this model. Thus, for the ignorable nonresponse model, combining (2) and (5), the joint prior density is

$$g_1(\mathbf{p}, \boldsymbol{\pi}) \propto 1, \quad p_{jk} \geq 0, \quad \sum_{j=1}^r \sum_{k=1}^c p_{jk} = 1, \quad \pi_s \geq 0, \quad \sum_{s=1}^4 \pi_s = 1, \quad (6)$$

which is proper.

Finally, combining the likelihood function in (4) with the joint prior density in (6) via Bayes' theorem, the joint posterior density of the parameters $\boldsymbol{\pi}, \mathbf{p}$ and $\mathbf{y}_{(1)}$ is

$$\boldsymbol{\pi}(\mathbf{p}, \boldsymbol{\pi}, \mathbf{y}_{(1)} | \mathbf{y}_1) \propto \left[\prod_{s=1}^4 \pi_s^{y_{s-}} \right] \left[\prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!} \right]. \quad (7)$$

A posteriori \mathbf{p} and $\boldsymbol{\pi}$ are independent. Inference about $\boldsymbol{\pi}$ is easy because $\boldsymbol{\pi} | \mathbf{y}_1, \mathbf{y}_{(1)} \sim \text{Dirichlet}(y_{1-} + 1, \dots, y_{4-} + 1)$, which is independent of $\mathbf{y}_{(1)}$. Inference about \mathbf{p} can be obtained using a simple Gibbs sampler because, letting $q_{jk}^{(1)} = p_{jk} / \sum_{k=1}^c p_{jk'}$ and $q_{jk}^{(2)} = p_{jk} / \sum_{j=1}^r p_{j'k}$, the conditional probabilities are

$$\mathbf{p} | \mathbf{y} \sim \text{Dirichlet}(y_{11} + 1, \dots, y_{rc} + 1),$$

$$y_{2j} | \mathbf{p}, u_j, \mathbf{y}_{(2)} \stackrel{\text{ind}}{\sim} \text{Multinomial}(u_j, \mathbf{q}_j^{(1)}), \quad j=1, \dots, r,$$

$$y_{3k} | \mathbf{p}, v_k, \mathbf{y}_{(3)} \stackrel{\text{ind}}{\sim} \text{Multinomial}(v_k, \mathbf{q}_k^{(2)}), \quad k=1, \dots, c,$$

$$\mathbf{y}_4 \mid \mathbf{p}, \mathbf{w}, \mathbf{y}_{(4)} \sim \text{Multinomial}(\mathbf{w}, \mathbf{p}). \quad (8)$$

Clearly, the parameters \mathbf{p} and $\boldsymbol{\pi}$ are identifiable and estimable. Also, note that \mathbf{y}_4 in (8) is a latent variable and that it does not contribute to inference about \mathbf{p} . Rather it assists in the computation by providing a simple Gibbs sampler. However, we note that information in \mathbf{y}_4 , via \mathbf{w} , is important under a nonignorable nonresponse model.

3.3 Nonignorable Nonresponse Model

For nonignorable missing data we take

$$\mathbf{J}_\ell \mid \{I_{j\ell} = 1, I_{j'\ell} = 0, j \neq j', k \neq k', \boldsymbol{\pi}_{jk}\} \stackrel{\text{iid}}{\sim} \text{Multinomial}\{1, \boldsymbol{\pi}_{jk}\}. \quad (9)$$

Assumption (9) specifies that the probabilities an individual belongs to one of the four tables depend on the two characteristics (*i.e.*, row and column classifications) of the individual. In this manner we incorporate the assumption that the missing data is nonignorable. This is an extension of the model in Nandram, Han and Choi (2002). One can also have $\boldsymbol{\pi}_j$ or $\boldsymbol{\pi}_k$ instead of $\boldsymbol{\pi}_{jk}$; the methodology is similar.

Next, we need the likelihood function. Here the augmented likelihood function for $\mathbf{p}, \boldsymbol{\pi}, \mathbf{y}_{(1)} \mid \mathbf{y}_1$ is

$$g(\mathbf{p}, \boldsymbol{\pi}, \mathbf{y}_{(1)} \mid \mathbf{y}_1, \mathbf{n}_0, \mathbf{u}, \mathbf{v}, \mathbf{w}) \propto \left[\prod_{s,j,k} \frac{\pi_{sjk}^{y_{sjk}}}{y_{sjk}!} \right] \left[\prod_{j,k} p_{jk}^{y_{jk}} \right], \quad (10)$$

subject to $\sum_{j=1}^r \sum_{k=1}^c y_{1jk} = n_0$, $\sum_{k=1}^c y_{2jk} = u_j$, $j = 1, \dots, r$, $\sum_{j=1}^r y_{3jk} = v_k$, $k = 1, \dots, c$, and $\sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w$.

Observe that in (10) the parameters p_{jk} and π_{sjk} are not identifiable. Clearly, to estimate p_{jk} one needs to know y_{jk} , but only the y_{1jk} are known. Also, to estimate π_{sjk} one needs to know y_{sjk} , $s = 2, 3, 4$. Thus, y_{sjk} , $s = 2, 3, 4$ are also not identifiable. Putting very informative proper priors on the π_{sjk} will help, but this is not a practical solution. If an ignorable model (*i.e.*, $\pi_{sjk} = \pi_s$) is used, then all the parameters can be identified. Therefore, a sensible solution is to attempt to link the π_{jk} over (j, k) using a common feature. If the π_{jk} come from a common distribution with “known” parameters, we would be able to estimate them. That is, we must attempt to “borrow strength” as in small area estimation. This permits estimation of $\mathbf{y}_{(1)}$ which, in turn, will facilitate estimation of the p_{jk} and π_{sjk} .

For the π_{jk} we “center” the nonignorable nonresponse model on the ignorable nonresponse model. Specifically, we assume that

$$\begin{aligned} \pi_{jk} \mid \boldsymbol{\mu}, \boldsymbol{\tau} &\stackrel{\text{iid}}{\sim} \text{Dirichlet}(\mu_1 \boldsymbol{\tau}, \mu_2 \boldsymbol{\tau}, \mu_3 \boldsymbol{\tau}, \mu_4 \boldsymbol{\tau}), \\ \pi_{sjk} &\geq 0, \sum_{s=1}^4 \pi_{sjk} = 1, \end{aligned} \quad (11)$$

$j = 1, \dots, r, k = 1, \dots, c$. In (11) the parameter $\boldsymbol{\tau}$ tells us about the closeness of the nonignorable nonresponse model to the ignorable nonresponse model. For example, if $\boldsymbol{\tau}$ is small, the π_{jk} will be very different, and if $\boldsymbol{\tau}$ is large, the π_{jk} will be very similar. Thus, inference may be sensitive to the choice of $\boldsymbol{\tau}$, and one has to be careful in choosing $\boldsymbol{\tau}$. In the absence of any information about nonignorability, it is natural to choose a prior density for $\boldsymbol{\tau}$ so that the nonignorable nonresponse model generalizes the ignorable nonresponse model. This generalization is attained because as $\boldsymbol{\tau}$ goes to infinity, the π_{jk} converge to the same value over (j, k) (not component-wise), the ignorable nonresponse model. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are not identifiable because the π_{jk} are not. Thus, it is impossible to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ without any information; a natural way to proceed is to attempt to use some of the data already observed.

Specifically, a priori we take $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ to be independent with

$$\begin{aligned} p(\boldsymbol{\mu}) &= 1, \mu_s \geq 0, s = 1, 2, 3, 4, \\ \sum_{s=1}^4 \mu_s &= 1, \boldsymbol{\tau} \sim \text{Gamma}(\alpha_0, \beta_0), \boldsymbol{\tau} \geq 0, \end{aligned} \quad (12)$$

where α_0 and β_0 are to be specified; without any information about α_0 and β_0 one needs to use the data again. To help specify α_0 and β_0 for the nonignorable nonresponse model, we have used the ignorable nonresponse model. The prior on $\boldsymbol{\tau}$ adds extra variation, thereby permitting some degree of nonignorability (see section 3.5). Note again that if $\boldsymbol{\tau}$ is very large (*i.e.*, $\alpha_0 \gg \beta_0$), this nonignorable nonresponse model degenerates into the ignorable nonresponse model. Thus, an issue of how sensitive inference is to this specification arises. Of course, one can choose other distributions for $\boldsymbol{\tau}$ in (12) (*e.g.*, lognormal distribution), but this is really not the key issue.

Combining (2), (11) and (12), the joint prior density of $\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\mu}$ and $\boldsymbol{\tau}$ is

$$\pi(\mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}) \propto \left\{ \prod_{j=1}^r \prod_{k=1}^c \frac{\prod_{s=1}^4 \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu} \boldsymbol{\tau})} \right\} \tau^{\alpha_0 - 1} e^{-\beta_0 \boldsymbol{\tau}}. \quad (13)$$

Note again that (13) is a proper prior density. Finally, combining the likelihood function in (10) with the joint prior density in (13) via Bayes' theorem, the joint posterior density of the parameters $\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\tau}$ and the latent variables $\mathbf{y}_{(1)}$ is

$$\begin{aligned} \pi(\mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{y}_{(1)} \mid \mathbf{y}_1) &\propto \left[\prod_{s,j,k} \frac{(\pi_{sjk} p_{jk})^{y_{sjk}}}{y_{sjk}!} \right] \\ &\times \left\{ \prod_{j,k} \frac{\prod_{s=1}^4 \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu} \boldsymbol{\tau})} \right\} \tau^{\alpha_0 - 1} e^{-\beta_0 \boldsymbol{\tau}}. \end{aligned} \quad (14)$$

In Appendix A we show how to fit the nonignorable nonresponse model to obtain the appropriate inference using the Gibbs sampler.

3.4 Bayes Factor: Tests of Association and Nonignorability

We construct a test for the association between BMD and FI. This test is an assessment of the assumption that $p_{jk} = q_{1j}q_{2k}$, $j = 1, \dots, r$, $k = 1, \dots, c$, and $\sum_{j=1}^r q_{1j} = 1$ and $\sum_{k=1}^c q_{2k} = 1$. We use the Bayes factor, the ratio of the marginal likelihoods under two scenarios (*e.g.*, association versus no association). Note that we observe $y_{(1)}$, but $y_{(1)}$ is a set of latent variables. So each marginal likelihood is simply the probability that y_1 is the observed value of Y_1 , which we denote by $p(y_1)$.

We set

$$C = \left\{ \begin{array}{l} y_{(1)} : \sum_{k=1}^c y_{2jk} = u_j, j = 1, \dots, r; \\ \sum_{j=1}^r y_{3jk} = v_k, k = 1, \dots, c; \sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w \end{array} \right\}.$$

Then, letting $d = 3!n!(rc-1)!$ and $e = 3!n!(r-1)!(c-1)!$, the marginal likelihood for the ignorable (IG) nonresponse model is

$$p_{IG}(y_1) = \left\{ \begin{array}{l} d \sum_{y_{(1)} \in C} \iint \prod_{s,j,k} \{(\pi_s p_{jk})^{y_{sjk}} / y_{sjk}!\} d\pi dp, \\ \text{association} \\ e \sum_{y_{(1)} \in C} \iint \prod_{s,j,k} \{(\pi_s q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}!\} d\pi dq_1 dq_2, \\ \text{no association,} \end{array} \right. \quad (15)$$

and letting $\Omega_a = (\pi, \mu, \tau)$ and $\Omega_{na} = (q_1, q_2, \pi, \mu, \tau)$, the marginal likelihood for the nonignorable (NIG) nonresponse model is

$$p_{NIG}(y_1) = \left\{ \begin{array}{l} d \sum_{y_{(1)} \in C} \int_{\Omega_a} \prod_{s,j,k} \{(\pi_{sjk} p_{jk})^{y_{sjk}} / y_{sjk}!\} g(\pi, \mu, \tau) d\Omega_a, \\ \text{association} \\ e \sum_{y_{(1)} \in C} \int_{\Omega_{na}} \prod_{s,j,k} \{(\pi_{sjk} q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}!\} g(\pi, \mu, \tau) d\Omega_{na}, \\ \text{no association,} \end{array} \right. \quad (16)$$

where

$$g(\pi, \mu, \tau) = \frac{\beta_0^{\alpha_0} \tau^{\alpha_0-1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} \prod_{j=1}^r \prod_{k=1}^c \left\{ \frac{\prod_{s=1}^4 \pi_{sjk}^{\mu_s \tau-1}}{D(\mu \tau)} \right\}. \quad (17)$$

The summation in the set C is computationally intensive because there are numerous points $y_{(1)} \in C$ (*i.e.*, we need to

sum over all of them). We avoid this problem by first summing over C analytically and the rest is obtained using Monte Carlo integration.

For the ignorable model it is easy to show that

$$p_{IG}(y_1) = \left\{ \begin{array}{l} a = \frac{3!n!}{n+1} \frac{(rc-1)!}{(n+rc-1)!}, \\ \text{association} \\ b = \frac{3!n!}{n+1} \frac{(r-1)!(c-1)!}{(n+r-1)!(n+c-1)!} \frac{\prod_j y_{1j}! \prod_k y_{1k}!}{\prod_j \prod_k y_{1jk}!}, \\ \text{no association,} \end{array} \right. \quad (18)$$

where n is the total number of individuals in the entire table. We describe how to estimate $p_{NIG}(y_1)$ in Appendix B.

However, we note that a test for ignorability or non-ignorability is tenuous because we assume that there is no information about ignorability or nonignorability. Yet, our nonignorable nonresponse model is a generalization of our ignorable nonresponse model. We believe that the test about association under the ignorable nonresponse model or nonignorable nonresponse model is reliable.

Finally, we note that the Bayes factor may be sensitive to prior specifications, especially when there are not enough data to estimate the parameters under test; see Sinharay and Stern (2002) for an interesting discussion on nested models. We have studied sensitivity of the Bayes factor with respect to the specification of α_0 and β_0 in (17); see section 3.5 and Table 6. This is useful because it is an important prior in our nonignorable nonresponse model. However, the main comparison is a test for no association under the ignorable nonresponse model and the nonignorable nonresponse model separately. The parameter τ only enters the non-ignorable nonresponse model, and τ has the same prior under association and no association.

3.5 Specification of α_0 and β_0

The specification of the hyperparameters α_0 and β_0 in $\tau \sim \text{Gamma}(\alpha_0, \beta_0)$ is a key issue in our method; see (12). This is important because we use this technique to robustify the ignorable nonresponse model; a sensitivity analysis is performed later. Note that $E(\tau) = \alpha_0 / \beta_0$; thus if $\alpha_0 \gg \beta_0$, the nonignorable nonresponse model will be similar to the ignorable nonresponse model. Suppose we can observe a random sample $\tau^{(1)}, \dots, \tau^{(M)}$ from $\text{Gamma}(\alpha_0, \beta_0)$. Then, we can use a simple method (*e.g.*, the method of moments) to estimate α_0 and β_0 .

How can we obtain a sample to fit $\text{Gamma}(\alpha_0, \beta_0)$? The Gibbs sampler in (8) for the ignorable nonresponse model gives imputed values for the missing cell counts. We have imputed the missing cell counts M times, $M = 1,000$;

let $n_{1jk}^{(h)} \equiv y_{1jk}$ and $n_{sjk}^{(h)}$, $s = 2, 3, 4$, $h = 1, \dots, M$ denote the missing cell counts. Then, for each h we fit the nonignorable nonresponse model without the prior specification in (12),

$$(n_{111}^{(h)}, \dots, n_{1rc}^{(h)}, \dots, n_{411}^{(h)}, \dots, n_{4rc}^{(h)}) | \pi, p \\ \sim \text{Multinomial}\{n, (\pi_{111}p_{11}, \dots, \pi_{4rc}p_{rc})\},$$

$$p \sim \text{Dirichlet}(\mathbf{1}), \text{ and } \pi_{jk} \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\alpha)$$

where $\alpha_s = \mu_s \tau$, $s = 1, 2, 3, 4$.

After integrating out p and π_{jk} , we get the likelihood function,

$$\prod_{j=1}^r \prod_{k=1}^c \left[\frac{\Gamma\left(\sum_{s=1}^4 \alpha_s\right)}{\Gamma\left(\sum_{s=1}^4 (\alpha_s + n_{sjk}^{(h)})\right)} \prod_{s=1}^4 \frac{\Gamma(\alpha_s + n_{sjk}^{(h)})}{\Gamma(\alpha_s)} \right], \\ \alpha_s > 0, s = 1, 2, 3, 4. \quad (19)$$

Using the Nelder-Mead algorithm to maximize the likelihood function in (19) over $\alpha_s > 0$, $s = 1, 2, 3, 4$, at the h^{th} iterate, we obtain the maximum likelihood estimators $\hat{\alpha}^{(h)}$, $h = 1, \dots, M$. Now letting $\tau^{(h)} = \sum_{s=1}^4 \hat{\alpha}_s^{(h)}$, we view $\tau^{(h)}$, $h = 1, \dots, M$ as a random sample from Gamma(α_0, β_0).

Finally, using the method of moments, we fit Gamma(α_0, β_0) to the “data”, $\tau^{(h)}$, $h = 1, \dots, M$, to get $\alpha_0 = a^2/b$ and $\beta_0 = a/b$, where $a = M^{-1} \sum_{h=1}^M \tau^{(h)}$ and $b = (M-1)^{-1} \sum_{h=1}^M (\tau^{(h)} - a)^2$. Thus, we have constructed a data-dependent prior distribution for τ . Our procedure gives $\alpha_0 = 125$, $\beta_0 = 0.35$ (i.e., τ has mean 357 and standard deviation 31.9). In section 4 we discuss sensitivity to this choice.

4. Data and Empirical Analysis

We apply our methodology to the data in the 3×3 categorical table in Table 1. After we present results associated with the observed data and a sensitivity analysis, we describe a simulation study to assess the difference between the ignorable and the nonignorable nonresponse models.

4.1 Data Analysis

See Table 2 for a comparison of the ignorable nonresponse model and the nonignorable nonresponse model. We have also included the numerical standard error (NSE) which is a measure of how well the numerical results can be reproduced; we have used the batch-means method to compute it. Thus, one would be comfortable with small NSE's relative to the Monte Carlo estimates or the posterior means. For both models the NSE's are small with relatively

larger values for the nonignorable nonresponse model (both near zero any way), indicating that the computations are repeatable. The posterior means (PM) are very similar for the two models. The posterior standard deviations (PSD) are larger for the nonignorable model, making the 95% credible intervals wider. Virtually all the 95% credible intervals under the ignorable nonresponse model are contained by those of the nonignorable nonresponse model.

Table 2

Comparison of the Posterior Means (PM), Posterior Standard Deviations (PSD), Numerical Standard Errors (NSE), and 95% Credible Intervals (CI) for p from the Ignorable and Nonignorable Nonresponse Models

Cell	\hat{p}	PM	PSD	NSE	CI
(a) Ignorable Model					
(1, 1)	0.337	0.330	0.005	0.001	(0.321, 0.339)
(1, 2)	0.157	0.142	0.003	0.001	(0.136, 0.147)
(1, 3)	0.154	0.168	0.004	0.001	(0.162, 0.175)
(2, 1)	0.141	0.142	0.004	0.001	(0.134, 0.148)
(2, 2)	0.071	0.066	0.002	0.001	(0.061, 0.070)
(2, 3)	0.063	0.071	0.003	0.001	(0.066, 0.078)
(3, 1)	0.050	0.053	0.003	0.001	(0.048, 0.059)
(3, 2)	0.016	0.016	0.001	0.000	(0.013, 0.019)
(3, 3)	0.010	0.012	0.002	0.000	(0.009, 0.015)
(b) Nonignorable Model					
(1, 1)	0.337	0.321	0.020	0.009	(0.278, 0.355)
(1, 2)	0.157	0.143	0.008	0.003	(0.126, 0.158)
(1, 3)	0.154	0.173	0.014	0.007	(0.140, 0.196)
(2, 1)	0.141	0.139	0.019	0.009	(0.109, 0.182)
(2, 2)	0.071	0.069	0.007	0.003	(0.056, 0.085)
(2, 3)	0.063	0.071	0.013	0.006	(0.053, 0.102)
(3, 1)	0.050	0.052	0.008	0.002	(0.040, 0.070)
(3, 2)	0.016	0.019	0.003	0.001	(0.014, 0.026)
(3, 3)	0.010	0.013	0.003	0.001	(0.009, 0.020)

Note: The ignorable nonresponse model has $\pi_{sjk} = \pi_s$, $s = 1, 2, 3, 4$, $j = 1, 2, 3$, $k = 1, 2, 3$. The observed value of p based on the complete data is \hat{p} .

In Table 3 we have also compared the estimation of π_s in the ignorable nonresponse model with π_{sjk} in the nonignorable nonresponse model. For the nonignorable nonresponse model we present the range of the posterior means (PM) for the nine cells of each s , $s = 1, 2, 3, 4$. This indicates the extent of the nonignorability. The PM's of π_s are within the range of the π_{sjk} , and as expected, the PSD's are larger for the nonignorable model. For example, over the nine cells the π_{1jk} vary from 0.388 to 0.656, and these two numbers differ significantly from 0.615, showing some degree of nonignorability. Thus, there is some difference between the ignorable and the nonignorable nonresponse models.

In Table 4 we have presented the logarithms of the Bayes factors for testing the goodness of fit of the ignorable nonresponse model and the nonignorable nonresponse model. There is “strong” evidence that the ignorable nonresponse model fits better than the nonignorable nonresponse model for these data (Kass and Raftery 1995). While the ignorable nonresponse model provides “strong” evidence for no association, the evidence from the nonignorable nonresponse model is “positive” as stated by Kass and Raftery (1995).

Thus, again there is a difference between the ignorable and the nonignorable nonresponse models. However, the NSE of 1.80 tends to nullify such differences. Our conclusion is that there is strong evidence to suggest no association between BMD and FI.

Table 3

Comparison of the Posterior Means (PM) and Posterior Standard Deviations (PSD) for π_{sjk} from the Ignorable and Nonignorable Nonresponse Models

	Ignorable	Nonignorable	
π_1	0.615 (0.009)	0.388 (0.078)	– 0.656 (0.044)
π_2	0.077 (0.005)	0.057 (0.017)	– 0.195 (0.068)
π_3	0.292 (0.008)	0.217 (0.041)	– 0.349 (0.053)
π_4	0.015 (0.002)	0.013 (0.005)	– 0.152 (0.055)

Note: PSD's are in parentheses. For the ignorable nonresponse model the parameters are π_1, π_2, π_3 and π_4 and for the nonignorable nonresponse model the parameters are π_{sjk} , $s = 1, 2, 3, 4$, $j = 1, 2, 3$, $k = 1, 2, 3$. Among the nine cells for each s we selected the smallest PM and the largest PM to form the range.

Table 4

Marginal Likelihoods and Bayes Factors for Testing Association Between BMD and FI Under the Ignorable and the Nonignorable Nonresponse Models

	Association	No association	Difference
Ignorable	–49.571	–46.173	–3.398
Nonignorable	–53.129	–50.132	–2.996
NSE	1.800	1.790	

Note: All entries (marginal likelihoods and their differences) are on the logarithmic scale. The Monte Carlo integration uses 50,000 iterations. The NSEs, numerical standard errors, are small relative to the marginal likelihoods.

We have considered the relation between BMD and FI when the osteopenia and osteoporosis levels are collapsed into one level. Under the ignorable nonresponse model the log Bayes factor is –2.77 (log marginal likelihoods: –32.82 and –29.05), and under the nonignorable nonresponse model the log Bayes factor is –4.52 (log marginal likelihoods: –34.25 and –4.52). Thus, the same conclusion is reached about no association between BMD and FI.

We have also separated out the data into two age groups: premenopausal (age at most 49 years old; young) and postmenopausal (age at least 50 years old; old). For the young group there were only 4 females with osteoporosis, and so we collapsed the females with osteopenia and osteoporosis. We fit both the ignorable and nonignorable nonresponse models to these data and got similar results. For the old group using the ignorable nonresponse model the log marginal likelihoods corresponding to no association and association are –43.01 and –38.91 giving a log Bayes factor of 4.10 for no association. Thus, there is strong evidence for no association between BMD and FI. For the young group using the ignorable nonresponse model the log marginal likelihoods corresponding to no association and association are –29.93 and –28.80 giving a log Bayes factor of 1.13 for no association. Thus, there is positive evidence for no association between BMD and FI for both age groups. Therefore, age is unlikely to play a role in the association of BMD and FI.

4.2 Sensitivity Analysis

We have studied the sensitivity of inference about the p_{jk} with respect to the prior distribution of τ . That is, we have taken $\tau \sim \text{Gamma}(\kappa\alpha_0, \beta_0)$, where κ is a sensitivity parameter that we have taken to be 1 in our analysis (note that $E(\tau) = \kappa\alpha_0 / \beta_0$).

Our procedure for the specification of α_0 and β_0 gives values of $\alpha_0 = 125$ and $\beta_0 = 0.35$; see section 3.5. Making κ bigger than 1 induces less changes in the posterior mean (PM) and posterior standard deviation (PSD) of the p_{jk} than for κ smaller than 1 because larger values of κ induces much smaller changes in the prior distribution of τ . In Table 5 we present PM's and PSD's of the p_{jk} for $\kappa = 0.25, 0.50, 1.00, 2.00, 4.00$. The PM's increase with κ and the PSD's decrease as κ increases from 0.25 to 4.00. Thus, there is some sensitivity to the specification of α_0 and β_0 , but the changes are small. For example, the PM's of p_{11} are 0.31, 0.32, 0.33 at $\kappa = 0.25, 1.00, 4.00$ and the PSD's at these values of κ are 0.04, 0.02, 0.01.

Table 5

Sensitivity of the Posterior Means (PM) and Posterior Standard Deviations (PSD) of the p_{jk} to Choices of κ in the Nonignorable Nonresponse Model

κ	0.25		0.50		1.00		2.00		4.00	
Cell	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
(1, 1)	306.93	36.09	315.01	25.81	321.81	19.95	325.37	14.55	326.16	10.46
(1, 2)	141.12	15.52	139.86	11.91	142.66	8.44	142.63	6.68	143.42	5.01
(1, 3)	161.68	25.80	167.83	18.77	173.40	13.77	176.20	8.44	175.78	6.71
(2, 1)	143.18	34.20	142.62	24.92	138.57	18.82	137.23	13.59	137.26	9.70
(2, 2)	68.46	13.12	71.06	10.09	68.44	7.48	68.79	5.72	68.11	4.45
(2, 3)	79.78	22.83	75.97	17.86	71.11	12.56	68.09	7.84	68.34	6.38
(3, 1)	59.97	21.60	53.50	12.12	52.14	7.76	50.97	5.29	51.41	4.35
(3, 2)	21.43	7.76	20.02	4.89	18.96	23.28	18.67	2.78	17.84	2.23
(3, 3)	17.45	10.38	14.12	4.28	12.93	2.99	12.05	2.34	11.69	1.99

Note: All entries must be multiplied by 10^{-3} . In the nonignorable nonresponse model $\pi_{sjk} \sim \text{Gamma}(\kappa\alpha_0, \beta_0)$, where κ is the sensitivity parameter and $\alpha_0 = 125$ and $\beta_0 = 0.35$.

We have also studied the sensitivity of the Bayes factors to choices of κ (see Table 6). First, the NSE's decrease with κ , but the change is small. Note that we have used 50,000 iterations in the Monte Carlo integration; this sample size is needed for the Monte Carlo estimates to stabilize. The log marginal likelihoods do not change too much with κ . Because the log Bayes factors are small, some changes are reflected in inference: At $\kappa = 0.25, 0.50, 4.00$ there is "strong" evidence for no association, but at $\kappa = 1.00, 2.00$ there is "positive" (borderline) evidence for no association. Overall, there is some degree of evidence for no association. Thus, it is interesting that one does not need to worry too much about the choice for (α_0, β_0) .

Table 6

Sensitivity of the Marginal Likelihoods and the Bayes Factor to Choices of κ in the Nonignorable Nonresponse Model

κ	Association		No Association		Bayes Factor
	ML	NSE	ML	NSE	
0.25	-53.37	1.90	-49.16	1.89	-4.21
0.50	-52.58	1.83	-49.49	1.82	-3.08
1.00	-52.58	1.80	-49.76	1.79	-2.82
2.00	-52.81	1.79	-49.83	1.78	-2.98
4.00	-52.95	1.78	-49.91	1.77	-3.04

Note: All entries are on the logarithm scale. In the nonignorable nonresponse model $\pi_{sjk} \sim \text{Gamma}(\kappa\alpha_0, \beta_0)$, where κ is the sensitivity parameter and $\alpha_0 = 125$ and $\beta_0 = 0.35$.

4.3 Simulation Study

We have performed a simulation study to further compare the ignorable and nonignorable nonresponse models. Our objective is to confirm differences that exist between the two models. In our situation a test based on the Bayes factor can confirm one or the other. With limited information about nonignorability (our current situation), it is sensible to fit an ignorable nonresponse model because all the parameters are identifiable in the ignorable nonresponse

model. Thus, we proceed by comparing the ignorable and nonignorable nonresponse models when data are generated from (a) the ignorable nonresponse model and (b) the nonignorable nonresponse model. This is a typical Bayesian analysis.

We obtained the posterior means of the p_{jk} and the π_{sjk} , denoted by \tilde{p}_{jk} and $\tilde{\pi}_{sjk}$ respectively, after the nonignorable nonresponse model is fit to the observed data. For the ignorable model we took $\tilde{\pi}_s = \sum_{j=1}^r \sum_{k=1}^c \tilde{\pi}_{sjk} / rc$, $s = 1, 2, 3, 4$. We obtained the cell counts for the ignorable model by drawing from

$$(y_{111}, \dots, y_{1rc}, \dots, y_{411}, \dots, y_{4rc}) | \tilde{\pi}, \tilde{p} \\ \sim \text{Multinomial}(n, (\tilde{\pi}_1 \tilde{p}_{11}, \dots, \tilde{\pi}_4 \tilde{p}_{rc}))$$

and for the nonignorable model by drawing from

$$(y_{111}, \dots, y_{1rc}, \dots, y_{411}, \dots, y_{4rc}) | \tilde{\pi}, \tilde{p} \\ \sim \text{Multinomial}(n, (\tilde{\pi}_{111} \tilde{p}_{11}, \dots, \tilde{\pi}_{4rc} \tilde{p}_{rc})),$$

where $n = 2,998$, the total number of individuals in the original data set (see Table 1). We have generated 1,000 datasets from each of the ignorable and nonignorable nonresponse models. Then, we fit the ignorable and nonignorable nonresponse models to each dataset in exactly the same manner for the observed data in Table 1, and we computed the posterior means (PM) and the posterior standard deviations (PSD) for the p_{jk} . In Table 7 we present the averages of the PM's and PSD's over the 1,000 datasets. The second column (labeled \hat{p}) has the posterior mean of p_{jk} for the observed data under the nonignorable nonresponse model (see Table 2b).

For (a) in Table 7 the PM's are very close to the \hat{p}_{jk} for the ignorable nonresponse model, but not so close when the nonignorable nonresponse model is fit. It is noticeable that

Table 7

Comparison of the Ignorable and Nonignorable Nonresponse Models Via the Simulated Data and the Posterior Means (PM) and Posterior Standard Deviations (PSD) of the p_{jk}

Cell	Simulated	Ignorable (a)				Nonignorable (b)			
	Fitted	Ignorable		Nonignorable		Ignorable		Nonignorable	
	\hat{p}	PM	PSD	PM	PSD	PM	PSD	PM	PSD
(1, 1)	321.81	320.73	5.72	307.42	11.30	332.02	5.10	324.44	10.60
(1, 2)	142.66	142.96	4.24	146.44	7.34	141.81	3.30	143.44	5.43
(1, 3)	173.40	172.59	4.42	173.49	7.62	168.66	4.14	174.10	7.04
(2, 1)	138.57	138.82	4.81	135.32	9.82	143.63	4.52	139.20	9.74
(2, 2)	68.44	68.44	3.55	72.01	6.02	64.51	2.91	68.20	4.76
(2, 3)	71.11	71.41	3.65	75.00	6.30	70.85	3.76	69.63	6.58
(3, 1)	52.14	52.17	3.11	53.03	4.95	53.08	3.04	52.44	4.70
(3, 2)	18.96	19.35	2.08	21.65	2.98	15.08	1.72	17.32	2.48
(3, 3)	12.93	13.54	1.78	15.64	2.55	10.95	1.85	11.20	2.18

Note: Data are simulated from the ignorable nonresponse model in (a) or the nonignorable nonresponse model in (b), and both the ignorable and nonignorable nonresponse models are fit. We have generated 1,000 datasets, and we fit both the ignorable and nonignorable nonresponse models to each simulated dataset. The PM's and PSD's are averages over the 1,000 datasets and \hat{p} is the posterior mean for the observed data which we used to generate the data sets. All entries must be multiplied by 10^{-3} .

the PSD's under the nonignorable nonresponse model are about twice as large as those under the ignorable nonresponse model. For (b) in Table 7 the PM's for the nonignorable nonresponse model are closer to the \hat{p}_{jk} than those from the ignorable nonresponse model. However, in both cases the PSD's for the nonignorable nonresponse model are about twice those from the ignorable nonresponse model. For example, in Table 7 for the (1, 1) cell as compared with 0.322 for \hat{p} , in (a) the ignorable (nonignorable) model gives a PM of 0.321 (0.307), but in (b) the ignorable (nonignorable) model gives a PM of 0.332 (0.324) for other examples. Thus, the two models are indeed different for estimating p .

We have also considered estimating the proportion P of simulated datasets in which the ignorable nonresponse model performs better than the nonignorable nonresponse model. It is expensive to compute the marginal likelihood under the nonignorable nonresponse model. We note again that it takes 50,000 iterations for the Monte Carlo estimate to stabilize; this is an enormous task for the simulation study because we need to calculate the marginal likelihoods for 1,000 datasets. Thus, we use a simple procedure to compare the two models, and we expect that this procedure would give a conclusion similar to a power calculation.

Specifically, we compute $\Delta^{(h)} = n \sum_{j=1}^r \sum_{k=1}^c (\hat{p}_{jk} - PM_{jk}^{(h)})^2 / PM_{jk}^{(h)}$, where $PM_{jk}^{(h)}$ is the posterior mean of p_{jk} corresponding to the h^{th} dataset. We denote $\Delta^{(h)}$ by $\Delta_{\text{IG}}^{(h)}$ for the ignorable nonresponse model and $\Delta_{\text{NIG}}^{(h)}$ for the nonignorable nonresponse model. An estimator of P , \hat{P} , is obtained by counting the number of the 1,000 experiments in which $\Delta_{\text{IG}}^{(h)} > \Delta_{\text{NIG}}^{(h)}$. For the data generated from the ignorable nonresponse model, \hat{P} is 0.236 with a standard error of 0.013. For the data generated from the nonignorable nonresponse model, \hat{P} is 0.920 with a standard error of 0.009. Thus, if the ignorable nonresponse model is expected to hold, about 24% of the time the nonignorable nonresponse model will beat it, and if the nonignorable nonresponse model is expected to hold, only about $(1 - 0.920)100\% \approx 8\%$ of the time the ignorable nonresponse model will beat it. Thus, there are latent differences between these two models. The nonignorable nonresponse model does capture some degree of nonignorability, and it robustifies the ignorable nonresponse model. We believe that this is a reasonable comparison between the ignorable and the nonignorable nonresponse models.

5. Concluding Remarks

There are two key methodological developments in this paper. Specifically, we have shown that (a) it is possible to analyze multinomial data from $r \times c$ categorical tables

when there are both item and unit nonresponses, and the nonresponse mechanism may be nonignorable; and (b) by using the Bayes factor (ratio of the marginal likelihoods of two models), we can test for association between the two categories. Essentially, we have assumed that there is no information about nonignorability, all design features are suppressed and we have taken a conservative ground.

For the 3×3 categorical data of BMD and FI, we have shown how to estimate the cell probabilities accurately. For the complete cases, the Bayes factor shows "strong" evidence for no association between BMD and FI. For all the data, our Bayes factor shows that the evidence for no association is "strong" under the ignorable nonresponse model, and is "positive" under the nonignorable nonresponse model. Thus, there is virtually no difference between the two scenarios: data from only the complete cases are used and all the data are used. Also, based on the Bayes factor and our simulation study, while there are differences between the ignorable nonresponse model and the nonignorable nonresponse models, such differences are small. There are differences for inference about the proportions of individuals in various BMD-FI levels; the posterior means are similar but the posterior standard deviations under the nonignorable nonresponse model are larger than those under the ignorable nonresponse model.

Our simulation study supports two properties (subtle differences) of our models. First, the estimates of the cell probabilities from the ignorable (nonignorable) nonresponse model are closer to the true values when the ignorable (nonignorable) nonresponse model is expected to hold, but in either case the estimates from the nonignorable nonresponse model have about twice the standard deviations from the ignorable nonresponse model. Second, if the ignorable (nonignorable) nonresponse model is expected to hold, it can be beaten by the nonignorable (ignorable) nonresponse model. This happens a significantly larger proportion of time when the ignorable nonresponse model is expected to hold. Thus, there are differences between these models. We suggest fitting both models, and compute the Bayes factor to decide which one to use. We do not recommend using these models when there are appropriate covariates and/or prior information to explain nonignorability.

In future research one can attempt to reduce the number of parameters in the nonignorable nonresponse model to further reduce the effects of nonignorability. For example, it may be possible to consider representing the data in two categorical tables as follows. The three supplemental tables are collapsed into a single supplemental table with its j^{th} row having at least u_j individuals, and its k^{th} column having at least v_k individuals; the total number of individuals in this supplemental table is $w + \sum_{j=1}^r u_j + \sum_{k=1}^c v_k$;

see section 3.1 for notations. Finally, we note that a full analysis of data from a complex survey requires an input of information (covariates and prior information) about non-ignorability, sampling weights and clustering effects as well.

Appendix A

Fitting the Nonignorable Nonresponse Model

We show how to use the Gibbs sampler to make inference about the parameters in (14). The conditional posterior density of p is

$$p | y \sim \text{Dirichlet}(y_{11} + 1, \dots, y_{rc} + 1) \quad (\text{A.1})$$

and the conditional posterior density of π_{jk} is

$$\pi_{jk} | \{\mu, \tau, y\} \stackrel{\text{ind}}{\sim} \text{Dirichlet} \left(\begin{array}{l} y_{1jk} + \mu_{1\tau}, y_{2jk} \\ + \mu_{2\tau}, y_{3jk} + \mu_{3\tau}, y_{4jk} + \mu_{4\tau} \end{array} \right) \quad (\text{A.2})$$

with independence over $j = 1, \dots, r, k = 1, \dots, c$.

We need the conditional posterior probability mass functions of $y_s, s = 2, 3, 4$ given $y_{(s)}, p, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c$. From (14) it is clear that the $y_s, s = 2, 3, 4$ are conditionally independent multinomial random vectors. Specifically,

$$\begin{aligned} y_{2j} | \{y_1, p, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c\} \\ \stackrel{\text{ind}}{\sim} \text{Multinomial}(u_j, q_j^{(2)}), j = 1, \dots, r, \\ y_{3k} | \{y_1, p, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c\} \\ \stackrel{\text{ind}}{\sim} \text{Multinomial}(v_k, q_k^{(3)}), k = 1, \dots, c, \\ y_4 | \{y_1, p, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c\} \\ \sim \text{Multinomial}(w, q^{(4)}), \end{aligned} \quad (\text{A.3})$$

where $q_{jk}^{(2)} = \pi_{2jk} p_{jk} / \sum_{k=1}^c \pi_{2jk'} p_{jk'}, k = 1, \dots, c, q_{jk}^{(3)} = \pi_{3jk} p_{jk} / \sum_{j=1}^r \pi_{3jk'} p_{jk'}, j = 1, \dots, r$, and $q_{jk}^{(4)} = \pi_{4jk} p_{jk} / \sum_{j=1}^r \sum_{k=1}^c \pi_{4jk'} p_{jk'}, j = 1, \dots, r, k = 1, \dots, c$.

Next, we consider the hyper-parameters. Letting $\delta_s = \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}$, the joint conditional posterior density of μ, τ is

$$p(\mu, \tau | \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c) \propto \left\{ \prod_{s=1}^4 \delta_s^{\mu_s \tau} \right\} / \{D(\mu, \tau)\}^{\tau c} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}, \quad (\text{A.4})$$

where $\sum_{s=1}^4 \mu_s = 1, \mu_s \geq 0, s = 1, 2, 3, 4, \tau > 0$.

We use the grid method to get samples from the conditional posterior density of $p(\mu | \tau, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c)$ and $p(\tau | \mu, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c)$. After transforming τ to $\phi/(1 - \phi)$, the parameters now live on $(0, 1)$ with appropriate constraints, making the grid procedure convenient. We use 50 intervals of equal widths (obtained by experimentation) to draw μ and ϕ , and a random deviate for τ is $\phi/(1 - \phi)$.

The Gibbs sampler is executed by drawing a random deviate from each of the conditional posterior “densities”, (A.1), (A.2), (A.3), and (A.4) in turn, iterating the entire procedure until convergence. This is an example of the griddy Gibbs sampler (Ritter and Tanner 1992).

Appendix B

Estimation of $p_{\text{NIG}}(y_1)$ in (16)

Letting n_m denote the number of incomplete cases (*i.e.*, $n = n_0 + n_m$), one can also show that for the model with association $p_{\text{NIG}}(y_1) = a((n+1)!/(n_0!n_m!))A$ and for the model with no association $p_{\text{NIG}}(y_1) = b((n+1)!/(n_0!n_m!))B$, where a and b are given in (18),

$$\begin{aligned} A = & \int_{\Omega_a} \left\{ \prod_{j,k} \pi_{1jk}^{y_{1jk}} \right\} \left\{ \sum_{s=2}^4 \sum_{j,k} \pi_{sjk} p_{jk} \right\}^{n_m} \left\{ \frac{\prod_{j,k} p_{jk}^{y_{1jk}}}{D(y_{111} + 1, \dots, y_{1rc} + 1)} \right\} \\ & \times \prod_{j,k} \left\{ \frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\mu, \tau)} \right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_a, \\ B = & \int_{\Omega_{na}} \left\{ \prod_{j,k} \pi_{1jk}^{y_{1jk}} \right\} \left\{ \sum_{s=2}^4 \sum_{j,k} \pi_{sjk} q_{1j} q_{2k} \right\}^{n_m} \\ & \frac{\prod_j q_{1j}^{y_{1j}}}{D(y_{111} + 1, \dots, y_{1rc} + 1)} \times \frac{\prod_k q_{2k}^{y_{1k}}}{D(y_{111} + 1, \dots, y_{1rc} + 1)} \\ & \prod_{j,k} \left\{ \frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\mu, \tau)} \right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_{na}. \end{aligned} \quad (\text{B.1})$$

Note that $0 < A, B < 1$ gives a useful diagnostic check on the computation.

We show how to compute A in (B.1) using Monte Carlo integration; the procedure to compute B is similar. We prefer the simpler procedure based on Monte Carlo integration with an importance function (Nandram and Kim 2002) rather than the method based on a continuation of the Gibbs sampler (Chib and Jeliazkov 2001).

For A we choose the importance function

$$\begin{aligned} \pi_{im}(\Omega_a) = & \frac{\prod_{j,k} p_{jk}^{y_{1jk}}}{D(y_{111} + 1, \dots, y_{1rc} + 1)} \prod_{j,k} \left[\frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\mu, \tau)} \right] \\ & \frac{\prod_{s=1}^4 \mu_s^{\mu_s \tau - 1}}{D(\tilde{\mu}, \tilde{\tau})} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} \end{aligned}$$

where $\tilde{\mu}_s$ and $\tilde{\tau}$ are estimates obtained using a Gibbs output. We obtain a sample from $\pi_{im}(\Omega_a)$ by drawing $\tau \sim \text{Gamma}(\alpha_0, \beta_0)$, $\mu \sim \text{Dirichlet}(\tilde{\mu}, \tilde{\tau})$, $\pi_{jk} | \mu, \tau \sim \text{Dirichlet}(\mu, \tau)$ and $p | y_1 \stackrel{\text{ind}}{\sim} \text{Dirichlet}(y_{111} + 1, \dots, y_{1rc} + 1)$.

Then, letting $w_h = \sum_j \sum_k y_{1jk} \log \pi_{1jk}^{(h)} + n_m \log [\sum_{s=2}^4 \sum_j \pi_{sjk}^{(h)} p_{jk}^{(h)}] - \sum_{s=1}^4 (\bar{\mu}_s \bar{\tau} - 1) + \log \mu_s^{(h)} + \log (D(\bar{\mu} \bar{\tau}))$, $h = 1, \dots, M$, an estimator of A is $\hat{A} = M^{-1} \sum_{h=1}^M e^{\omega_h}$. The numerical standard error (NSE) of $\log(\hat{A})$ can be approximated. For letting $\bar{\omega} = M^{-1} \sum_{h=1}^M \omega_h$ and $S^2 = (M-1)^{-1} \sum_{h=1}^M (\omega_h - \bar{\omega})^2$, we have $\text{Var}(\hat{A}) \approx e^{2\bar{\omega}} S^2 / M$, $\text{Var}(\log(\hat{A})) \approx (\text{Var}(\hat{A}) / e^{2\bar{\omega}}) \approx S^2 / M$, and the NSE is S / \sqrt{M} approximately. We start with $M = 10,000$ independent samples from the importance function, and increasing M until convergence which occurs about $M = 50,000$.

Acknowledgements

This was a part of the work done during the academic year 2003/2004 when Balgobin Nandram was on sabbatical as a Research Scientist at the National Center for Health Statistics, Hyattsville, Maryland. We are grateful to the Associate Editor and the two referees for their informative comments, and the three opportunities to revise the manuscript.

References

- Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.
- Chib, S., and Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96, 270-281.
- Cohen, G., and Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents. *Journal of Official Statistics*, 18, 13-23.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45-97.
- Farahmand, B.Y., Persson, P.G., Michaelsson, K., Baron, J.A., Parker, M.G. and Ljunghall, S. (2000). Socioeconomic status, marital status and hip fracture risk: a population-based case control study. *Osteoporosis International*, 11, 803-808.
- Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society, Series B*, 60, 57-70.
- Ganry, O., Baudoin, C. and Fardellone, P. (2000). Effect of alcohol intake on bone mineral density in elderly women: The EPIDOS Study. *American Journal of Epidemiology*, 151, 8, 773-780.
- Kass, R., and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Lauderdale, D.S., and Rathouz, P.J. (2003). Does bone mineralization reflect economic conditions? An examination using a national US sample. *Economics and Human Biology*, 1, 91-104.
- Little, R.J. (2003). Bayesian Approach to Sample Survey Inference. In *Analysis of Survey Data*, (Eds. R.L. Chambers and C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.
- Little, R.J.A., and Rubin D.B. (2002). *Statistical Analysis with Missing Data*. Edition, New York: John Wiley & Sons, Inc.
- Looker, A.C., Orwoll, E.S., Johnston, C.C., Lindsay, R.L., Wahner, H.W., Dunn, W., Calvo, M.S. and Harris, T.B. (1997). Prevalence of low femoral bone density in older U.S. adults from NHANES III. *Journal of Bone and Mineral Research*, 12, 1761-1768.
- Looker, A.C., Wahner, H.W., Dunn, W.L., Calvo, M.S., Harris, R.R., Heyse, S.P., Johnston, C.C. and Lindsay, R. (1998). Updated data on proximal femur bone mineral levels of us adults. *Osteoporosis International*, 8, 468-489.
- Mirkin, B. (2001). Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, 55, 111-120.
- Nandram, B., and Choi, J.W. (2002 a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Nandram, B., and Choi, J.W. (2002 b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.
- Nandram, B., and Choi, J.W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the nhanes data. *Survey Methodology*, 31, 73-84.
- Nandram, B., Han, G. and Choi, J.W. (2002). A hierarchical Bayesian nonignorable nonresponse model for multinomial data from small areas. *Survey Methodology*, 28, 145-156.
- Nandram, B., and Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, 72, 319-340.
- Nandram, B., Liu, N., Choi, J.W. and Cox, L.H. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.
- Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- Ritter, C., and Tanner, M.A. (1992). The Gibbs stopper and the griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Rubin, D.B., Stern, H.S. and Vehovar, V. (1995). Handling "Don't know" survey responses: The case of the slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.
- Sinharay, S., and Stern, H.S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56, 196-201.
- Wang, H. (2001). *Two-way Contingency Tables with Marginally and Conditionally Imputed Nonrespondents*, Ph.D. Dissertation, Department of Statistics, University of Wisconsin-Madison.

On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment

Jean-François Beaumont¹

Abstract

Nonresponse weight adjustment is commonly used to compensate for unit nonresponse in surveys. Often, a nonresponse model is postulated and design weights are adjusted by the inverse of estimated response probabilities. Typical nonresponse models are conditional on a vector of fixed auxiliary variables that are observed for every sample unit, such as variables used to construct the sampling design. In this note, we consider using data collection process variables as potential auxiliary variables. An example is the number of attempts to contact a sample unit. In our treatment, these auxiliary variables are taken to be random, even after conditioning on the selected sample, since they could change if the data collection process were repeated for a given sample. We show that this randomness introduces no bias and no additional variance component in the estimates of population totals when the nonresponse model is properly specified. Moreover, when nonresponse depends on the variables of interest, we argue that the use of data collection process variables is likely to reduce the nonresponse bias if they provide information about the variables of interest not already included in the nonresponse model and if they are associated with nonresponse. As a result, data collection process variables may well be beneficial to handle unit nonresponse. This is briefly illustrated using the Canadian Labour Force Survey.

Key Words: Nonresponse bias; Nonresponse model; Nonresponse variance; Number of attempts; Paradata; Response probability.

1. Introduction

Unit nonresponse is often handled in surveys by using a nonresponse weight adjustment method. The basic principle that is often chosen is to adjust the design weights by the inverse of estimated response probabilities (see, for example, Ekholm and Laaksonen 1991). These estimated response probabilities are obtained by postulating a model for the unknown nonresponse mechanism, which we call the nonresponse model. Key to reducing the nonresponse bias and variance as much as possible is to condition on a vector of auxiliary variables that are observed for every sample unit and that are good predictors of both nonresponse and the variables of interest (Little and Vartivarian 2005). Usually, the auxiliary variables are treated as being fixed both unconditionally and conditionally on the selected sample.

In this note, we consider using Data Collection Process (DCP) variables as potential auxiliary variables to be included in the nonresponse model. An example is the number of attempts to contact a sample unit. Such type of data is sometimes called paradata (see Couper and Lyberg 2005 for a recent reference on paradata) and has been used to deal with unit nonresponse by Holt and Elliott (1991), among others. In our treatment, contrary to Holt and Elliott (1991), DCP variables are taken to be random, even after conditioning on the selected sample, since they could

change if the data collection process were repeated for a given sample.

DCP variables may be particularly useful in cross-sectional surveys where the auxiliary variables available to handle unit nonresponse are often limited to variables used to construct the sampling design. Although such design variables are not useless, they are often neither very good predictors of nonresponse nor the variables of interest. The additional information from data collection process may be welcome in these cases. In longitudinal surveys, there is a wealth of potential auxiliary variables to deal with wave nonresponse. DCP information may thus not have the same importance to compensate for wave nonresponse than the importance it has to compensate for unit nonresponse in cross-sectional surveys. However, we have yet to study this in any depth. It may turn out that, at change points, DCP variables may matter greatly.

In section 2, we introduce notation and theory concerning the effect of using random auxiliary variables in the nonresponse model when estimating population totals. This issue of the randomness of DCP auxiliary variables was raised and debated at Statistics Canada's Advisory Committee on Statistical Methods after the paper by Alavi and Beaumont (2004) was presented. The goal of section 2 is thus to shed some light on this issue. The use of DCP variables to adjust design weights for nonresponse is briefly illustrated in section 3, using the Canadian Labour Force

1. Jean-François Beaumont, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.
E-mail: jean-francois.beaumont@statcan.ca.

Survey (CLFS). The last section, section 4, contains a brief summary of the paper.

2. Theory

Let us assume that we are interested in estimating the population total $t_y = \sum_{k \in U} y_k$ of a variable of interest y for a certain fixed population U of size N . From this population, a random sample s of size n is selected according to a probability sampling design $p(s|\mathbf{D})$, where \mathbf{D} is a N -row matrix containing \mathbf{d}'_k in its k^{th} row and \mathbf{d} is the vector of design variables. Let also assume that, in the absence of nonresponse, we would use the Horvitz-Thompson estimator $\hat{t}_y = \sum_{k \in s} w_k y_k$, where $w_k = 1/\pi_k$ is the design weight of unit k and $\pi_k = P(k \in s)$ is its selection probability.

Usually, due to a number of reasons, unit nonresponse occurs so that the variable y is only observed for a subset s_r of s , the respondents. Along with s_r , a random vector \mathbf{z} of DCP variables is also observed for every sample unit according to a joint mechanism $\#q(\mathbf{Z}_s, s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$. As mentioned in the introduction, the number of attempts to contact a sample unit is an example of a DCP variable. The vector \mathbf{z} of DCP variables and the set of respondents s_r are random after conditioning on the selected sample since these quantities would likely take different values if the data collection process were repeated for a given sample. The quantity \mathbf{Z}_s is a n -row matrix containing \mathbf{z}'_k in its k^{th} row, \mathbf{Y} is a N -element vector containing y_k in its k^{th} element and \mathbf{X} is a N -row matrix containing \mathbf{x}'_k in its k^{th} row. The vector \mathbf{x} is a vector of additional fixed auxiliary variables. For instance, these auxiliary variables could come from an administrative file or, in a longitudinal survey, they could be the variables of interest observed at the previous wave. As a result, the vector \mathbf{x} may not be available for nonsample units. Table 1 summarizes the availability of the different types of variables for the respondents, nonrespondents and nonsample units.

Table 1
Availability of Variables

	y	\mathbf{z}	\mathbf{x}	\mathbf{d}
Respondents: s_r	YES	YES	YES	YES
Nonrespondents: $s - s_r$	NO	YES	YES	YES
Nonsample units: $U - s$	NO	NO*	YES**	YES

* The vector \mathbf{z} is not even defined for nonsample units.
** The vector \mathbf{x} may not always be available for nonsample units.

The joint mechanism $\#q(\mathbf{Z}_s, s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$ can be factorized into two distinct random mechanisms: i) $\#(\mathbf{Z}_s | s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$ and ii) $q(s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s)$. The

former is called the DCP mechanism while the latter is called the nonresponse mechanism. This factorization will be useful later to obtain properties of our nonresponse-weight-adjusted estimator defined in equation (2.2) below. We assume that

$$q(s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s) = q(s_r | s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s), \quad (2.1)$$

where \mathbf{D}_s and \mathbf{X}_s are the sample portions of \mathbf{D} and \mathbf{X} respectively. This assumption implies that the nonresponse mechanism is independent of (or unconfounded with) \mathbf{Y} , after conditioning on $s, \mathbf{D}_s, \mathbf{X}_s$ and \mathbf{Z}_s , and that the data are missing at random. However, we make no explicit simplifying assumption about the DCP mechanism so that it may well depend on \mathbf{Y} , even after conditioning on s, \mathbf{D} and \mathbf{X} .

To compensate for unit nonresponse, we consider the nonresponse-weight-adjusted estimator

$$\hat{t}_y^{\text{NWA}} = \sum_{k \in s} \frac{w_k}{p_k(\hat{\mathbf{a}})} y_k, \quad (2.2)$$

where $p_k(\mathbf{a}) = P(k \in s_r | s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s; \mathbf{a})$ is the conditional response probability for a unit $k \in s$ and $\hat{\mathbf{a}}$ is an estimator of the vector of unknown nonresponse model parameters \mathbf{a} . Note that a nonresponse model is a set of assumptions about the unknown nonresponse mechanism $q(s_r | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s)$; one of them being assumption (2.1). We assume that $\hat{\mathbf{a}}$ is implicitly defined by the equation $\mathbf{U}_1(\hat{\mathbf{a}}) = \mathbf{0}$, where $\mathbf{U}_1(\cdot)$ is a vector of q -unbiased estimating functions for \mathbf{a} ; that is, $\mathbf{E}_q\{\mathbf{U}_1(\mathbf{a}) | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s\} = \mathbf{0}$. Therefore, $\mathbf{U}_1(\cdot)$ is also $p\#q$ -unbiased for \mathbf{a} . In the remaining of the paper, we remove everywhere the conditioning on \mathbf{Y}, \mathbf{D} and \mathbf{X} when taking expectations and variances since these vectors are always treated as being fixed. For instance, we will write $\mathbf{E}_q\{\mathbf{U}_1(\mathbf{a}) | s, \mathbf{Z}_s\} = \mathbf{0}$ instead of $\mathbf{E}_q\{\mathbf{U}_1(\mathbf{a}) | s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s\} = \mathbf{0}$. This will simplify considerably the notation.

Note that the nonresponse-weight-adjusted estimator (2.2) is implicitly defined by the equation

$$U_2(\hat{\mathbf{a}}, \hat{t}_y^{\text{NWA}}) = \hat{t}_y^{\text{NWA}} - \sum_{k \in s} \frac{w_k}{p_k(\hat{\mathbf{a}})} y_k = 0. \quad (2.3)$$

If the nonresponse model is correctly specified and, in particular, if assumption (2.1) is satisfied, then the estimating function $U_2(\cdot, \cdot)$ is $p\#q$ -unbiased for t_y ; that is, $\mathbf{E}_{p\#q}\{U_2(\mathbf{a}, t_y)\} = 0$. To make assumption (2.1) as plausible as possible, it is important that the nonresponse model be conditional on design, auxiliary and DCP variables that are well correlated with y , provided that these variables are also associated with nonresponse. This recommendation should be useful to control the magnitude of the nonresponse bias, which may be unavoidable in real surveys.

This is also in line with the recommendation given in Little and Vartivarian (2005). Therefore, if DCP variables contain information about y above the information already contained in \mathbf{d} and \mathbf{x} , then the use of DCP variables may be useful to reduce the nonresponse bias if they are associated with nonresponse.

Now, let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', t_y)'$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}', \hat{t}_y^{NWA})'$ and $\mathbf{U}(\tilde{\boldsymbol{\theta}}) = \{\mathbf{U}_1'(\tilde{\boldsymbol{\alpha}}), \mathbf{U}_2'(\tilde{\boldsymbol{\alpha}}, \tilde{t}_y)\}'$, for some vector $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}', \tilde{t}_y)'$. As noted above, $\hat{\boldsymbol{\theta}}$ is implicitly defined by the equation $\mathbf{U}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and the estimating function $\mathbf{U}(\cdot)$ is $p\#q$ -unbiased for $\boldsymbol{\theta}$ since $\mathbf{E}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$. Using a first-order Taylor approximation (see Binder 1983), we have $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} - \{\mathbf{H}(\boldsymbol{\theta})\}^{-1} \mathbf{U}(\boldsymbol{\theta})$, where $\mathbf{H}(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{p\#q}\{\partial \mathbf{U}(\tilde{\boldsymbol{\theta}}) / \partial \tilde{\boldsymbol{\theta}}'\}$. The matrix $\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}$ is thus given by

$$\{\mathbf{H}(\boldsymbol{\theta})\}^{-1} = \begin{pmatrix} \{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & \mathbf{0} \\ -\mathbf{H}_{21}(\boldsymbol{\theta})\{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & 1 \end{pmatrix}, \quad (2.4)$$

where $\mathbf{H}_{i1}(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{p\#q}\{\partial \mathbf{U}_i(\tilde{\boldsymbol{\theta}}) / (\partial \tilde{\boldsymbol{\alpha}}')\}$, for $i=1, 2$. Using conditions similar to those of Binder (1983), $\hat{\boldsymbol{\theta}}$ is asymptotically normal and asymptotically $p\#q$ -unbiased for $\boldsymbol{\theta}$. As a result, \hat{t}_y^{NWA} is asymptotically normal and asymptotically $p\#q$ -unbiased for t_y . Therefore, using DCP variables in the nonresponse model does not introduce any bias in the nonresponse-weight-adjusted estimator \hat{t}_y^{NWA} provided that the nonresponse model (specification of $q(s_r | s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s)$ and assumption 2.1) holds. Also, if the true unknown nonresponse mechanism depends on the sample portion of \mathbf{Y} , \mathbf{Y}_s , after conditioning on s , \mathbf{D}_s and \mathbf{X}_s , then conditioning on a vector \mathbf{z} of DCP variables is likely to reduce the nonresponse bias if the DCP mechanism depends on \mathbf{Y}_s , after conditioning on s , \mathbf{D}_s and \mathbf{X}_s , which means that the DCP variables contain information about y not already contained in \mathbf{d} and \mathbf{x} .

Continuing our Taylor linearization, and using the fact that

$$\begin{aligned} \mathbf{V}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\} &= \mathbf{V}_p \mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) | s\} \\ &\quad + \mathbf{E}_p \mathbf{V}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) | s, \mathbf{Z}_s\} \\ &\quad + \mathbf{E}_{p\#} \mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) | s, \mathbf{Z}_s\}, \end{aligned}$$

the $p\#q$ -variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, $\mathbf{V}_{p\#q}(\hat{\boldsymbol{\theta}})$, is approximated by

$$\begin{aligned} \dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}}) &= \{\mathbf{H}(\boldsymbol{\theta})\}^{-1} \mathbf{V}_p \mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) | s\} \{\mathbf{H}'(\boldsymbol{\theta})\}^{-1} \\ &\quad + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1} \mathbf{E}_p \mathbf{V}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) | s, \mathbf{Z}_s\} \{\mathbf{H}'(\boldsymbol{\theta})\}^{-1} \\ &\quad + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1} \mathbf{E}_{p\#} \mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) | s, \mathbf{Z}_s\} \{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}. \quad (2.5) \end{aligned}$$

The first term on the right-hand side of equation (2.5) is called the sampling variance of $\hat{\boldsymbol{\theta}}$, the second term is called the DCP variance of $\hat{\boldsymbol{\theta}}$ and the third term is called the nonresponse variance of $\hat{\boldsymbol{\theta}}$. The variance $\mathbf{V}_{p\#q}(\hat{t}_y^{NWA})$ is

approximated by the value in the last row and in the last column of equation (2.5). Using expression (2.4) and the fact that $\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) | s, \mathbf{Z}_s\} = (\mathbf{0}', t_y - \hat{t}_y)'$, the approximate variance (2.5) reduces to

$$\begin{aligned} \dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}}) &= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_p(\hat{t}_y) \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &\quad + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1} \mathbf{E}_{p\#} \mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) | s, \mathbf{Z}_s\} \{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}. \quad (2.6) \end{aligned}$$

The second matrix on the right-hand side of equation (2.6) corresponds to the DCP variance of $\hat{\boldsymbol{\theta}}$ and contains 0 for all its elements. Therefore, using random auxiliary (DCP) variables in the nonresponse model does not introduce any additional term of variance, as opposed to using only fixed auxiliary variables, when the nonresponse model is properly specified. Since DCP variables are likely to reduce the nonresponse bias if they are associated with y , then it seems beneficial to take advantage of them when handling unit nonresponse through a weight adjustment. Also, as pointed out by Little and Vartivarian (2005), adding auxiliary variables in the nonresponse model that are associated with y tends to reduce the nonresponse variance. The mean squared error can therefore be reduced on both counts.

A more detailed expression for the nonresponse variance term in equation (2.6) as well as a sampling and a nonresponse variance estimator can be obtained similarly as in Beaumont (2005). Beaumont (2005) also discusses the effect of estimating the nonresponse model parameters on the variance of an estimator of a population total.

3. The Example of the Canadian Labour Force Survey

The goal of this example is not to provide every detail of the analysis that was conducted on the Canadian Labour Force Survey (CLFS) data but simply to describe some issues related to the choice of the nonresponse model and to the estimation of response probabilities. With these points in mind, we then go on to discuss the main conclusions that were reached. Greater detail about the results of the investigations in the CLFS, implementation of the new method and a comparison with the previous method can be found in Alavi and Beaumont (2004).

The CLFS is a monthly survey with a stratified multi-stage sampling design (Gambino, Singh, Dufour, Kennedy and Lindeyer 1998). The information used to construct the sampling design and to draw a sample of dwellings is essentially geographic. The sample is divided into six representative rotation groups and each sampled dwelling stays in the sample for six consecutive months. One rotation group contains dwellings for which the members are interviewed

for the first time; another rotation group contains dwellings for which the members are interviewed for the second time and so on. Thus, for five rotation groups out of six, the sampled dwellings are common from one month to the next. Computer-assisted interviews are used to collect the survey information for every person in the selected households. With computer-assisted interviews, a large amount of DCP information is obtained for both responding and nonresponding households.

A logistic nonresponse model has been considered to model the unknown nonresponse mechanism $q(s, l, \mathbf{D}_s, \mathbf{Z}_s)$. With this model, the unknown response probability for household k is expressed as $p_k(\boldsymbol{\alpha}) = \{1 + \exp(-\boldsymbol{\alpha}'(\mathbf{z}\mathbf{d})_k)\}^{-1}$ and sampled households are assumed to respond independently of one another. The vector $\mathbf{z}\mathbf{d}$ is a vector that contains DCP variables \mathbf{z} , fixed design variables \mathbf{d} as well as interactions between these two types of variables. No additional vector \mathbf{x} of auxiliary variables was available. Two DCP variables were used: the number of attempts to contact a sampled household, which was divided into five categories, and the time of the last attempt, which was also divided into five categories. The design variables used were mainly geographic and also included the rotation group indicator. Due to potential interviewer and clustering effects, the above model may not be entirely realistic. It was used for its simplicity and because it appeared reasonable and an improvement over the previous method. Also, the estimated response probabilities resulting from this model were used only to provide a score and were not used directly to adjust design weights, as described below in this section.

The unknown vector $\boldsymbol{\alpha}$ was estimated by the maximum likelihood method using the q -unbiased estimating function

$$\mathbf{U}_1(\boldsymbol{\alpha}) = \sum_{k \in s} \{r_k - p_k(\boldsymbol{\alpha})\}(\mathbf{z}\mathbf{d})_k, \quad (3.1)$$

where $r_k = 1$, if $k \in s$, and $r_k = 0$, otherwise. Note that a design-weighted estimating function was not considered. This follows the practice recommended in Little and Vartivarian (2003) and can be justified by noting that the interest is in modelling the nonresponse mechanism only for sampled households $k \in s$ (not for the whole population) and that this mechanism is conditional on s . Also, the DCP variables are not even defined outside the sample. The use of design weights does thus not make sense in this context and increases the variance of $\hat{\boldsymbol{\alpha}}$ if the nonresponse model is correctly specified. Also, it is not clear that using a design-weighted estimating function would systematically bring robustness in this case. However, note that we do not ignore design information since it is included in the nonresponse model. This can be paralleled to the recommendation of including design information in imputation models (see, for example, Rubin 1996).

Stepwise logistic regression was performed for several months in order to determine appropriate design and DCP variables to be included in the final nonresponse model. In all months considered, the variable 'number of attempts' was the first to enter in the model and thus the most useful for explaining nonresponse. This variable was also correlated with the main variables of interest 'employment' and 'unemployment'. For instance, people belonging to respondent households with a large number of attempts, i.e. those that are difficult to reach, had a tendency to be more often employed (see Alavi and Beaumont 2004). Households with a large number of attempts had also a tendency to be nonrespondents. Therefore, it seems appropriate to give a larger weight adjustment to the responding households for which the number of attempts is large since their propensity to respond is lower and they are more likely to have characteristics similar to the nonrespondents.

The final nonresponse model chosen fit reasonably well the CLFS data in most months considered, according to the Hosmer-Lemeshow goodness-of-fit test. Nevertheless, the score method of Little (1986) was used to obtain some robustness against undetected model failures. The above logistic nonresponse model was first used to obtain an estimated response probability for every sampled household and then the sample was divided into about 50 homogeneous classes with respect to this estimated response probability using the clustering algorithm implemented in the procedure FASTCLUS of SAS. This large number of classes was possible given the large CLFS sample size. It was chosen so as to reduce the nonresponse bias not only at the population level but also for smaller domains. The nonresponse weight adjustment for a responding household k within a given class c was simply computed as the inverse of the unweighted response rate within class c . A threshold on the nonresponse weight adjustment was set to 2.5 to control the nonresponse variance of the nonresponse-weight-adjusted estimator. When needed, the application of this threshold was necessary only for a very small number of classes. These were the classes with the smallest estimated response probabilities. Without this threshold, nonresponse weight adjustments around 4 could occasionally be observed.

Another nonresponse model was considered in which the response probability for a household k is modelled as the product of the probability that household k be contacted, times the probability that this household respond, given it is contacted. The latter two probabilities were modelled separately. Although this model seems to be a better approximation of reality and gave slightly better results in the sense that it better explained nonresponse, the gains were not deemed sufficient to add this complexity in the nonresponse adjustment method. It may deserve further study.

4. Conclusion

An important contribution of this paper is that DCP information must be treated as being random when used in a nonresponse model. We then have shown that the use of such information to handle unit nonresponse through a weight adjustment does not introduce any bias and that there is no additional variance component in the estimates of population totals when the nonresponse model is properly specified. Moreover, we have argued that if DCP information is associated with the variables of interest and with nonresponse, then its use is likely to reduce the nonresponse bias when the nonresponse mechanism depends directly on the variables of interest. We have also illustrated through the CLFS example that such information can be useful for dealing with unit nonresponse in a major survey.

The full response estimator that we have considered is the Horvitz-Thompson estimator. Our conclusions would have remained the same had we used instead a generalized regression estimator. We have used the Horvitz-Thompson estimator for its simplicity and because it was sufficient to show our main point.

Acknowledgements

I would like to thank the members of Statistics Canada's Advisory Committee on Statistical Methods for raising issues with the application of the proposed method in the Labour Force Survey and, particularly, J.N.K. Rao and Chris Skinner for useful discussions following the presentation to the Committee. I would also like to sincerely thank the Associate Editor for his comments and suggestions. They were very useful and improved the clarity of the paper. Finally, I am much indebted to Asma Alavi and Cynthia Bocci of Statistics Canada for preparing computer programs used to analyse Labour Force Survey data.

References

- Alavi, A., and Beaumont, J.-F. (2004). Nonresponse adjustment plans for the Labour Force Survey. Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, May 2-3, 2004.
- Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B*, 67, 445-458.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Couper, M., and Lyberg, L. (2005). The use of paradata in survey research. *Bulletin of the International Statistical Institute* (to appear).
- Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 7, 325-337.
- Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindey, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue number 71-526.
- Holt, D., and Elliott, D. (1991). Methods of weighting for unit nonresponse. *The Statistician*, 40, 333-342.
- Little, R.J. (1986). Survey nonresponse adjustment for estimate of means. *International Statistical Review*, 54, 139-157.
- Little, R.J., and Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 22, 1589-1599.
- Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 161-168.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

On the Correlation Structure of Sample Units

Alfredo Bustos¹

Abstract

In this paper we make explicit some distributional properties of sample units, not usually found in the literature; in particular, their correlation structure and the fact that it does not depend on arbitrarily assigned population indices. Such properties are relevant to a number of estimation procedures, whose efficiency would benefit from making explicit reference to them.

Key Words: Census; Survey; Sampling; Sample units; Probability function; Mean; Covariance.

1. Introduction

In recent times, population and household censuses, as we know them, have become more difficult to perform for a number of reasons. Alternative ways of securing more frequent information for the production of local, state and national statistical results have been proposed. Continuous large national surveys, among them those known as rolling censuses, with large sample sizes and complex designs, are being considered.

However, in order to produce results at the local authority level the way a census does, different techniques for estimation as well as for validation and, in some cases, for imputation have to be developed and their efficiency improved. One way of achieving greater efficiency consists of taking into account all relevant information available. Of course, this includes the stochastic properties of sample units.

In what follows, beginning from basic principles, we derive a general explicit form for the probability function of an ordered sample. We also show how that function, as well as the inclusion probabilities, can be computed. Finally, we give a general form for the correlation matrix of sample units, which depends solely on inclusion probabilities, so that linear and maximum-likelihood estimation procedures can benefit from it.

2. The Basic Model

The basic model we start from represents the sequential random drawing of n units from a population U formed by N such units, and may be stated as follows. Let N and n be two positive constants such that $n \leq N$, and let V represent an $N \times n$ matrix, whose components are each distributed as Bernoulli random variables with, possibly, different parameters. Then,

$$V_{N \times n} = \begin{bmatrix} \vartheta_{11} & \vartheta_{12} & \vartheta_{13} & \cdots & \vartheta_{1n} \\ \vartheta_{21} & \vartheta_{22} & \vartheta_{23} & \cdots & \vartheta_{2n} \\ \vartheta_{31} & \vartheta_{32} & \vartheta_{33} & \cdots & \vartheta_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vartheta_{N1} & \vartheta_{N2} & \vartheta_{N3} & \cdots & \vartheta_{Nn} \end{bmatrix}. \quad (1.1)$$

Also part of the model is the restriction imposed on each column of V to add to one. In other words, we require that

$$\sum_{i=1}^N \vartheta_{ik} = 1, \text{ for } k = 1, \dots, n \quad (1.2)$$

be satisfied.

This is required because if the j^{th} draw results in population unit I being selected, then entry (I, j) takes the value of one while all other entries of column j are equal to zero. Note that this is equivalent to imposing a non-stochastic constraint on the behavior of all components of the i^{th} column of V , regardless of the sampling scheme. Therefore, entries belonging to the same column do not behave independently.

When sampling takes place with replacement (WR), the sum of the elements of the I^{th} row of the above matrix is distributed as a Binomial (n, p_I) since each column is distributed independently of other columns. On the other hand, when sampling takes place without replacement (WOR), the total of row I can take only two values: one, if the I^{th} unit is drawn at some stage, or zero, otherwise, bringing us back to the Bernoulli case.

Disjoint subsets of rows may be formed according to different criteria. For instance, when rows are grouped with regard to their spatial vicinity, one could speak about clusters or primary sampling units. When one or more statistical indicators form the basis for the groupings, the term strata is usually used.

1. Victor Alfredo Bustos y de la Tijera, Instituto Nacional de Estadística, Geografía e Informática, H. de Nacozari 2301, 20270, Aguascalientes, Ags., México. E-mail: alfredo.bustos@inegi.gob.mx.

Let us now define the inclusion probabilities as

$$\pi_I^{(k)} = P(\text{population unit } I \text{ in sample of size } k) \\ = 0 \text{ if } k = 0. \quad (2)$$

Note that $\pi_I^{(n)} = \pi_I$, commonly referred to as the inclusion probability for unit I .

Now let $\vartheta_{\cdot j}$ represent the j^{th} column and $\vartheta_{I \cdot}$ the I^{th} row of matrix V . Therefore, based on the following expression,

$$f(\vartheta_{\cdot 1}, \vartheta_{\cdot 2}, \vartheta_{\cdot 3}, \dots, \vartheta_{\cdot n}) = f(\vartheta_{\cdot 1})f(\vartheta_{\cdot 2} | \vartheta_{\cdot 1}) \\ f(\vartheta_{\cdot 3} | \vartheta_{\cdot 1}, \vartheta_{\cdot 2}) \dots f(\vartheta_{\cdot n} | \vartheta_{\cdot 1}, \dots, \vartheta_{\cdot n-1}) \quad (3)$$

we can write the joint probability function of the elements of V as:

$$f(\vartheta_{\cdot 1}, \vartheta_{\cdot 2}, \vartheta_{\cdot 3}, \dots, \vartheta_{\cdot n}) = \prod_{k=1}^n \left[\prod_{I=1}^N (\pi_I^{(k)} - \pi_I^{(k-1)})^{\vartheta_{I k}} \right] \\ = \prod_{k=1}^n \left[\prod_{I=1}^N (p_I^{(k)})^{\vartheta_{I k}} \right] \quad (4)$$

subject to

$$\sum_{I=1}^N \vartheta_{I k} = 1, k = 1, \dots, n \text{ and} \\ \sum_{k=1}^N \vartheta_{I k} \leq \begin{cases} 1, \text{ WOR} \\ n, \text{ WR} \end{cases} I = 1, \dots, N;$$

and here $p_I^{(k)}$, defined as $p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)})$, stands for the probability that population unit I is included in the sample at the k^{th} draw. The above function is useful for calculating the probability of any ordered sample of size n . Clearly, when the order of inclusion can be ignored, the probability of a given sample would be obtained by adding the $n!$ values obtained through (4).

3. The Implications of Sampling on the Stochastic Properties of Population Units

Consequently,

$$E(\vartheta_{I k}) = p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)}) \quad (5)$$

and therefore, we can write

$$E[V] = \begin{bmatrix} p_1^{(1)} & p_1^{(2)} & p_1^{(3)} & \dots & p_1^{(n)} \\ p_2^{(1)} & p_2^{(2)} & p_2^{(3)} & \dots & p_2^{(n)} \\ p_3^{(1)} & p_3^{(2)} & p_3^{(3)} & \dots & p_3^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_N^{(1)} & p_N^{(2)} & p_N^{(3)} & \dots & p_N^{(n)} \end{bmatrix}. \quad (6)$$

From here, step-by-step inclusion probabilities, in WOR sampling situations, may be recursively computed, as is shown in (7), below.

$$p_I^{(k)} = \begin{cases} p_I & \text{if } k = 1 \\ p_I^{(k-1)} \sum_{j \neq I}^N \frac{p_j^{(k-1)}}{1 - p_j^{(k-1)}} & \text{if } k > 1. \end{cases} \quad (7)$$

Note that (7) enables us to compute the desired probabilities at two different moments: first, when no draw has actually occurred, which explains why we average over the whole population, and secondly, when the result of the previous draw is known, at which time the probability of the J^{th} population unit, say, entering the sample equals one and all other probabilities for that draw are equal to zero. Hence, at least in theory, we can compute the inverse of the so called expansion factors or weights for one stage sampling, or stage by stage in multistage sampling. Clearly,

$$\pi_I^{(n)} = \sum_{k=1}^n p_I^{(k)}. \quad (8)$$

If we define the joint inclusion probabilities as

$$\pi_{IJ}^{(k)} = P \left(\begin{matrix} \text{population units } I \text{ and} \\ J \text{ in sample of size } k \end{matrix} \right), \quad (9)$$

then we have that they can also be computed as follows:

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^n p_j^{(k)} + p_J^{(j)} \sum_{k>j}^n p_I^{(k)} \right). \quad (10)$$

For example, in simple random sampling WR (SRS/WR), expressions (7), (8) and (10) result in (7.1), (8.1) and (10.1),

$$p_I^{(k)} = \frac{1}{N} \text{ when } k \geq 1 \quad (7.1)$$

$$\pi_I^{(n)} = \frac{n}{N} \quad (8.1)$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^n p_j^{(k)} + p_J^{(j)} \sum_{k>j}^n p_I^{(k)} \right) \\ = \sum_{j=1}^{n-1} \left(\frac{n-j}{N^2} + \frac{n-j}{N^2} \right) = \frac{n(n-1)}{N^2}. \quad (10.1)$$

While in SRS/WOR we get expressions (7.2), (8.2) and (10.2), instead.

$$p_I^{(k)} = \frac{1}{N} \text{ when } k \geq 1 \quad (7.2)$$

$$\pi_I^{(n)} = \frac{n}{N} \quad (8.2)$$

$$\begin{aligned} \pi_{IJ}^{(n)} &= \sum_{j=1}^{n-1} \left(p_I^{(j)} \sum_{k>j}^n p_J^{(k)} + p_J^{(j)} \sum_{k>j}^n p_I^{(k)} \right) \text{ where } J \neq I \\ &= \sum_{j=1}^{n-1} \left(\frac{n-j}{N(N-1)} + \frac{n-j}{N(N-1)} \right) = \frac{n(n-1)}{N(N-1)}. \end{aligned} \quad (10.2)$$

Let us now consider the row vectors $\underline{\vartheta}_{I_o}$. Then, for the covariance matrix between different rows, we get

$$\text{Cov}(\underline{\vartheta}_{I_o}, \underline{\vartheta}_{J_o}) = \begin{bmatrix} -p_I^{(1)} p_J^{(1)} & 0 & \cdots & 0 \\ 0 & -p_I^{(2)} p_J^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -p_I^{(n)} p_J^{(n)} \end{bmatrix}_{n \times n} \quad (11)$$

whenever I is different from J .

When sampling takes place WR, and therefore, $p_I^{(j)} = p_I \forall j=1, \dots, n$, the covariance matrix for the I^{th} row vector is given by

$$\text{Cov}(\underline{\vartheta}_{I_o}, \underline{\vartheta}_{I_o}) = \begin{bmatrix} p_I q_I & 0 & 0 & \cdots & 0 \\ 0 & p_I q_I & 0 & \cdots & 0 \\ 0 & 0 & p_I q_I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_I q_I \end{bmatrix}_{n \times n}. \quad (12.1)$$

In a WOR setting the above covariance matrix becomes

$$\text{Cov}(\underline{\vartheta}_{I_o}, \underline{\vartheta}_{I_o}) = \begin{bmatrix} p_I^{(1)}(1-p_I^{(1)}) & -p_I^{(1)} p_I^{(2)} & \cdots & -p_I^{(1)} p_I^{(n)} \\ -p_I^{(1)} p_I^{(2)} & p_I^{(2)}(1-p_I^{(2)}) & \cdots & -p_I^{(2)} p_I^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_I^{(1)} p_I^{(n)} & -p_I^{(2)} p_I^{(n)} & \cdots & p_I^{(n)}(1-p_I^{(n)}) \end{bmatrix}_{n \times n}. \quad (12.2)$$

Let $\underline{\vartheta}$ represent the N -dimensional vector which results from adding the columns of V . Clearly, the components of this vector may be expressed as the product of $\underline{\vartheta}_{I_o}$ by a vector whose components are all equal to one. In other words,

$$\underline{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ \vdots \\ \vartheta_N \end{pmatrix} = \begin{pmatrix} \underline{\vartheta}_{1_o}^T \underline{1} \\ \underline{\vartheta}_{2_o}^T \underline{1} \\ \underline{\vartheta}_{3_o}^T \underline{1} \\ \vdots \\ \underline{\vartheta}_{N_o}^T \underline{1} \end{pmatrix}. \quad (13)$$

Some distributional properties of these sums may be then obtained directly from those of the rows or the columns of matrix V .

For instance, their expected values are given as

$$\begin{aligned} E(\vartheta_I) &= E(\underline{\vartheta}_{I_o}^T \underline{1}) = E\left(\sum_{k=1}^n \vartheta_{I_k}\right) \\ &= \sum_{k=1}^n p_I^{(k)} = \pi_I^{(1)} + \sum_{k=2}^n (\pi_I^{(k)} - \pi_I^{(k-1)}) = \pi_I^{(n)}. \end{aligned} \quad (14)$$

From (1.2), we get the non-stochastic restriction:

$$\underline{1}' \underline{\vartheta} = \vartheta_1 + \vartheta_2 + \vartheta_3 + \dots + \vartheta_N = n. \quad (15)$$

From (14) and (15), well known propositions (16) and (17) follow immediately,

$$E(\underline{\vartheta}') = (\pi_1^{(n)}, \pi_2^{(n)}, \pi_3^{(n)}, \dots, \pi_N^{(n)}) \quad (16)$$

$$\pi_1^{(n)} + \pi_2^{(n)} + \pi_3^{(n)} + \dots + \pi_N^{(n)} = n. \quad (17)$$

For the second order moments, we get

$$\begin{aligned} \text{Cov}(\vartheta_I, \vartheta_J) &= \text{Cov}(\underline{1}' \underline{\vartheta}_{I_o}, \underline{1}' \underline{\vartheta}_{J_o}) \\ &= \underline{1}' \text{Cov}(\underline{\vartheta}_{I_o}, \underline{\vartheta}_{J_o}) \underline{1} = -\sum_{k=1}^n p_I^{(k)} p_J^{(k)} \\ &= \begin{cases} -np_I p_J & \text{WR} \\ (\pi_{IJ}^{(n)} - \pi_I^{(n)} \pi_J^{(n)}) & \text{WOR}, \end{cases} \end{aligned} \quad (18)$$

which clearly indicates that the covariance is never positive. In turn, the variances are given by

$$\begin{aligned} \text{Var}(\vartheta_I) &= \text{Var}(\underline{1}' \underline{\vartheta}_{I_o}) = \underline{1}' \text{Cov}(\underline{\vartheta}_{I_o}) \underline{1} \\ &= \begin{cases} np_I q_I & \text{WR} \\ \pi_I^{(n)} (1 - \pi_I^{(n)}) & \text{WOR}. \end{cases} \end{aligned} \quad (19)$$

Another important consequence of (15) has to do with the second order moments of the stochastic vector $\underline{\vartheta}$.

$$0 = \text{Var}(n) = \text{Var}(\underline{1}' \underline{\vartheta}) = \underline{1}' \text{Cov}(\underline{\vartheta}) \underline{1} = \underline{1}' C \underline{1}. \quad (20)$$

Clearly, the diagonal elements of matrix C , the covariance matrix of $\underline{\vartheta}$, are not all equal to zero. Therefore, randomly drawing a fixed-size simple introduces a dependency in the population units which results in non-null covariances implying that matrix C is singular. Otherwise, it is impossible for (20) to be satisfied.

As a matter of fact, it is possible to prove that the sum of any row (or column) of C must be equal to zero, which is a stronger statement. Given that the covariance between a random variable and a constant equals zero, we get

$$\begin{aligned}
0 &= \text{Cov}(\vartheta_I, n) = \text{Cov}(\vartheta_I, \vartheta_1 + \vartheta_2 + \vartheta_3 + \dots + \vartheta_N) \\
&= C_{I1} + C_{I2} + \dots + C_{IN} \\
&= \text{Var}(\vartheta_I) + \sum_{J \neq I} \text{Cov}(\vartheta_I, \vartheta_J). \quad (21)
\end{aligned}$$

We have thus proven that in WOR sampling (22.1) holds.

$$0 = \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}). \quad (22.1)$$

The same statement can be proven algebraically by noting that

$$\begin{aligned}
\sum_{J \neq I} \pi_{IJ}^{(n)} &= \pi_I^{(n)} \sum_{J \neq I} \pi_{J|I}^{(n)} \\
&= (n-1)\pi_I^{(n)},
\end{aligned}$$

which is obvious once we realize that the conditional probability involved represents the probability that population unit J enters a sample of size $n-1$ for which (19) also applies. Additionally, using (19) again, note that

$$\sum_{J \neq I} \pi_J^{(n)} = (n - \pi_I^{(n)}),$$

and therefore,

$$\begin{aligned}
0 &= \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}) \\
&= \pi_I^{(n)} - (\pi_I^{(n)})^2 + (n-1)\pi_I^{(n)} - \pi_I^{(n)}(n - \pi_I^{(n)}).
\end{aligned}$$

For WR sampling (21) implies:

$$\begin{aligned}
0 &= np_I q_I + \sum_{J \neq I} (n(n-1)p_I p_J - n^2 p_I p_J) \\
&= np_I q_I - np_I \sum_{J \neq I} p_J \quad (22.2)
\end{aligned}$$

which is immediately seen to apply.

In any case, the most important implication of the above results is that regardless of the sampling scheme, the correlation matrix of the population random variables $\vartheta_1, \vartheta_2, \vartheta_3, \dots, \vartheta_N$ is singular. For the practical situations described in the introduction, the most important implication of this fact lies mainly in the use made by many model fitting and estimation procedures of the inverse of the covariance matrix.

4. The First Two Moments of Sample Units

Once the first and second order moments of the vector ϑ have been established, we are in a position to determine the corresponding moments for sub-vectors of different sizes and whose components are randomly chosen, *i.e.*, the sample. To this end, let us define the random variables $\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \dots, \vartheta_{I_r}$, where r represents the number of different population units in the sample, and whose indices $I_k, 1 \leq k \leq r \leq n$, can take the value I with probability $\pi_I^{(n)}$. In other words, under the above conditions, we are in

the presence of a set of random variables whose indices are random themselves.

4.1 Mean and Variance for WR Sampling

For this case, the probability function of ϑ_{I_1} is given by

$$\begin{aligned}
P(\vartheta_{I_1} = x) &= \sum_{I=1}^N p_I P(\vartheta_I = x) \\
&= \sum_{I=1}^N p_I \binom{n}{x} p_I^x (1 - p_I)^{n-x}. \quad (23)
\end{aligned}$$

The first two moments may also be obtained via a conditional argument. The mean of its distribution is given by

$$E(\vartheta_{I_1}) = \sum_{I=1}^N p_I E(\vartheta_I) = \sum_{I=1}^N np_I p_I = n \sum_{I=1}^N p_I^2. \quad (24)$$

In turn, its variance is computed using the well known formula

$$V(\vartheta_{I_1}) = V_{I_1}[E(\vartheta_{I_1} | I_1)] + E_{I_1}[V(\vartheta_{I_1} | I_1)]. \quad (25)$$

In this case, we have

$$\begin{aligned}
E(\vartheta_{I_1} | I_1 = I) &= np_I \\
\text{and } V(\vartheta_{I_1} | I_1 = I) &= np_I(1 - p_I). \quad (26)
\end{aligned}$$

Hence,

$$\begin{aligned}
V_{I_1}[E(\vartheta_{I_1} | I_1)] &= V_{I_1}(np_{I_1}) \\
&= n^2 [E_{I_1}(p_{I_1}^2) - E_{I_1}^2(p_{I_1})], \\
E_{I_1}[V(\vartheta_{I_1} | I_1)] &= n E_{I_1}[p_{I_1}(1 - p_{I_1})] \\
&= n [E_{I_1}(p_{I_1}) - E_{I_1}(p_{I_1}^2)] \quad (27)
\end{aligned}$$

and therefore

$$\begin{aligned}
V(\vartheta_{I_1}) &= n [E_{I_1}(p_{I_1}) - E_{I_1}(p_{I_1}^2)] + n^2 [E_{I_1}(p_{I_1}^2) - E_{I_1}^2(p_{I_1})] \\
&= \sum_{I=1}^N np_I^2 \left(1 + (n-1)p_I - \sum_{J=1}^N np_J^2 \right). \quad (28)
\end{aligned}$$

For the case of SRS, (24) above results in

$$E(\vartheta_{I_1}) = \frac{1}{n} \sum_{I=1}^N (\pi_I^{(n)})^2 = \frac{1}{n} \sum_{I=1}^N \left(\frac{n}{N} \right)^2 = \left(\frac{n}{N} \right)^2 \frac{N}{n} = \frac{n}{N}.$$

While (28) yields

$$V(\vartheta_{I_1}) = \sum_{I=1}^N n \frac{1}{N^2} \left(1 + (n-1) \frac{1}{N} - \sum_{J=1}^N n \frac{1}{N^2} \right) = n \frac{1}{N} \left(1 - \frac{1}{N} \right).$$

4.2 Mean and Variance for WOR Sampling

For this case, the probability function of ϑ_{I_1} is given by

$$P(\vartheta_{I_1} = x) = \frac{1}{n} \sum_{I=1}^N \pi_I^{(n)} \sum_{k=1}^n (p_I^{(k)})^x (1 - p_I^{(k)})^{1-x} \quad (29)$$

and therefore

$$\begin{aligned} E(\vartheta_{I_j}) &= \frac{1}{n} \sum_{l=1}^N \pi_l^{(n)} E(\vartheta_l) \\ &= \frac{1}{n} \sum_{l=1}^N \pi_l^{(n)} \sum_{j=1}^n (p_l^{(j)}) = \frac{1}{n} \sum_{l=1}^N (\pi_l^{(n)})^2. \end{aligned} \quad (30)$$

Using (25) again, we note firstly that

$$E(\vartheta_{I_j} | I_j) = \pi_{I_j}^{(n)} \text{ and } V(\vartheta_{I_j} | I_j) = \pi_{I_j}^{(n)}(1 - \pi_{I_j}^{(n)})$$

from which we get

$$V[E(\vartheta_{I_j} | I_j)] = V(\pi_{I_j}^{(n)}) = E[(\pi_{I_j}^{(n)})^2] - [E(\pi_{I_j}^{(n)})]^2$$

and

$$E[V(\vartheta_{I_j} | I_j)] = E[\pi_{I_j}^{(n)}(1 - \pi_{I_j}^{(n)})] = E[(\pi_{I_j}^{(n)})] - [E(\pi_{I_j}^{(n)})]^2.$$

Hence, the variance is given by

$$\begin{aligned} V(\vartheta_{I_j}) &= E(\pi_{I_j}^{(n)}) - E^2(\pi_{I_j}^{(n)}) = E(\pi_{I_j}^{(n)})[1 - E(\pi_{I_j}^{(n)})] \\ &= \left(\frac{1}{n} \sum_{l=1}^N (\pi_l^{(n)})^2 \right) \left[1 - \left(\frac{1}{n} \sum_{l=1}^N (\pi_l^{(n)})^2 \right) \right]. \end{aligned} \quad (31)$$

Once again, in order to exemplify these results, let us turn to SRS. Expression (30) becomes

$$\begin{aligned} E(\vartheta_{I_j}) &= \frac{1}{n} \sum_{l=1}^N (\pi_l^{(n)})^2 \\ &= \frac{1}{n} \sum_{l=1}^N \left(\frac{n}{N} \right)^2 = \left(\frac{n}{N} \right)^2 \frac{N}{n} = \frac{n}{N}. \end{aligned} \quad (32)$$

Whereas (31) results in

$$\begin{aligned} V(\vartheta_{I_j}) &= \left(\frac{1}{n} \sum_{l=1}^N \left(\frac{n}{N} \right)^2 \right) \left[1 - \left(\frac{1}{n} \sum_{l=1}^N \left(\frac{n}{N} \right)^2 \right) \right] \\ &= \frac{n}{N} \left(1 - \frac{n}{N} \right). \end{aligned} \quad (33)$$

4.3 The Covariance Between Sample Units

In order to establish the covariance between different sample units we resort to a simple extension to (25),

$$\begin{aligned} \text{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \text{Cov}_{I_j, I_k} [E(\vartheta_{I_j} | I_j), E(\vartheta_{I_k} | I_k)] \\ &\quad + E_{I_j, I_k} [\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j, I_k)]. \end{aligned} \quad (34)$$

In this case, we have that

$$E(\vartheta_{I_j} | I_j = I) = \pi_I^{(n)} \quad (35)$$

and

$$E(\vartheta_{I_j}, \vartheta_{I_k} | I_j = I, I_k = J) = \pi_{IJ}^{(n)} \quad (36)$$

while the covariance between brackets on the right-hand side of (34) is easily seen to equal

$$\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j = I, I_k = J) = \pi_{IJ}^{(n)} - \pi_I^{(n)} \pi_J^{(n)}. \quad (37)$$

From (35) and (36), we obtain

$$\begin{aligned} \text{Cov}_{I_j, I_k} [E(\vartheta_{I_j} | I_j), E(\vartheta_{I_k} | I_k)] \\ = E_{I_j, I_k} (\pi_{I_j I_k}^{(n)} \pi_{I_k}^{(n)}) - E_{I_j} (\pi_{I_j}^{(n)}) E_{I_k} (\pi_{I_k}^{(n)}) \end{aligned} \quad (38)$$

whereas from (37) we get

$$\begin{aligned} E_{I_j, I_k} [\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j, I_k)] \\ = E_{I_j, I_k} (\pi_{I_j I_k}^{(n)}) - E_{I_j} (\pi_{I_j}^{(n)}) E_{I_k} (\pi_{I_k}^{(n)}). \end{aligned} \quad (39)$$

Finally, adding these last two expressions we arrive at the desired covariance

$$\begin{aligned} \text{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) \\ = E_{I_j, I_k} (\pi_{I_j I_k}^{(n)}) - [E_{I_j} (\pi_{I_j}^{(n)})][E_{I_k} (\pi_{I_k}^{(n)})] \\ = \frac{1}{n(n-1)} \sum_{l=1}^N \sum_{j \neq l}^N (\pi_{lj}^{(n)})^2 - \left(\frac{1}{n} \sum_{l=1}^N (\pi_l^{(n)})^2 \right)^2. \end{aligned} \quad (40)$$

In the SRS/WR (40) results in

$$\begin{aligned} \text{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \frac{1}{n(n-1)} \sum_{l=1}^N \sum_{j \neq l}^N \left(\frac{n(n-1)}{N^2} \right)^2 \\ &\quad - \left(\frac{1}{n} \sum_{l=1}^N \left(\frac{n}{N} \right)^2 \right)^2 \\ &= \frac{n(n-1)}{N^2} - \frac{n^2}{N^2} \\ &= -\frac{n}{N^2}, \end{aligned} \quad (41)$$

while for the WOR case the covariance can be seen to equal

$$\begin{aligned} \text{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \frac{1}{n(n-1)} \sum_{l=1}^N \sum_{j \neq l}^N \left(\frac{n(n-1)}{N(N-1)} \right)^2 \\ &\quad - \left(\frac{1}{n} \sum_{l=1}^N \left(\frac{n}{N} \right)^2 \right)^2 \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \\ &= -\left(\frac{n(N-n)}{N^2(N-1)} \right). \end{aligned} \quad (42)$$

It should be stressed that for SRS, regardless of whether it takes place with or without replacement, the correlation coefficients are given by

$$\text{Corr}(\vartheta_{I_j}, \vartheta_{I_k}) = \frac{-1}{(N-1)}, \quad (43)$$

independently of the sample size.

Furthermore, we have that, as the value of n approaches that of N in WOR sampling, both $\pi_I^{(n)}$ and $\pi_{II}^{(n)}$ approach one. In particular, when $n = N$, the values of expressions (31) and (40) become zero.

5. The Correlation Matrix for Sample Units

Once we realize that none of the expressions in (28), (31) and (40) depend on any of the arbitrary indices used to differentiate population units, it should become clear that the $r \times r$ correlation matrix for the random vector $\underline{\theta} = (\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \dots, \vartheta_{I_r})$, where $r \leq n$, may be written as:

$$\text{Corr}(\underline{\theta}) = R_r(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}. \quad (44)$$

It should be noted that the elements of $R_r(\rho)$ in (44) depend only on the inclusion probabilities which, for any sample size, may be fully computed from recursion (7), and

expressions (8) and (10). In other words, they do not depend on any unknown population parameters to be estimated nor on the values of the variables to be measured on the sample units.

6. Final Remarks

In theory, the efficiency of every estimation procedure will experience some gain whenever explicit allowance for the correlation between sample units is made. This would certainly be the case for linear as well as for some instances of maximum-likelihood estimation.

On the other hand, it should be emphasized that $R_n(\rho)$ may become singular as the sample size n approaches the population size N ; this is the case for SRS ($R_N(-1/(N-1))$) as well as for WOR sampling in general. Therefore, numerically, many estimation procedures which rely on the inverse or the determinant of R , rather than on the correlation matrix itself, may also benefit from replacing the simplifying assumption of independence between observations by a more realistic one of correlated observations whenever sample sizes are large relative to population sizes. Instances where this can happen are given by some stages in multi-stage sampling (*e.g.*, number of households in a block) and by large country-wide surveys.

Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling

Changbao Wu¹

Abstract

We present computational algorithms for the recently proposed pseudo empirical likelihood method for the analysis of complex survey data. Several key algorithms for computing the maximum pseudo empirical likelihood estimators and for constructing the pseudo empirical likelihood ratio confidence intervals are implemented using the popular statistical software R and S-PLUS. Major codes are written in the form of R/S-PLUS functions and therefore can directly be used for survey applications and/or simulation studies.

Key Words: Confidence interval; Bi-section algorithm; Empirical likelihood; Newton-Raphson procedure; Stratified sampling; Unequal probability sampling.

1. Introduction

One of the major challenges in applying advanced and often sophisticated statistical methods for real world surveys is the computational implementation of the method. Practical considerations often rule out the use of methods which are theoretically sound and attractive but are computationally formidable.

The empirical likelihood method first proposed by Owen (1988) is one of the major advances in statistics during the past fifteen years. In addition to its data driven and range respecting feature in estimation and testing, its non-parametric and discrete nature is particularly appealing for finite population problems. Indeed an early version of the method, the so-called scale-load estimators, was used in survey sampling by Hartley and Rao back in 1968. The more recent investigation of the method in survey sampling has resulted in a series of research papers and generated noticeable interests among survey statisticians to further explore the method. Wu and Rao (2004) contains a brief summary on the recent development of the pseudo empirical likelihood (PEL) method in survey sampling.

Progress on algorithmic development for the PEL method has also been made. A modified Newton-Raphson procedure for computing the maximum PEL estimators under non-stratified sampling was proposed by Chen, Sitter and Wu (2002). The procedure was further modified by Wu (2004a) to handle stratified sampling designs.

In this article we present computational algorithms for computing the maximum PEL estimators and for constructing the related PEL ratio confidence intervals for complex surveys under a unified framework, with particular interest in implementing those algorithms using R and S-PLUS. The software package R, a friendly programming

environment and compatible to the popular commercial statistical software S-PLUS, is attracting more and more users from the statistical community. What is advantageous about using R is that it is available free for research use and the package may be easily downloaded from the web. It is hoped that this article will bridge the current gap between theoretical developments and practical applications of the PEL method and will generate more research activities in this direction to make fully practical use of the PEL method a reality.

The algorithm for computing the maximum PEL estimator under non-stratified sampling and some notes on its implementation in R/S-PLUS are presented in section 2. The algorithm of Wu (2004a) for stratified sampling is discussed in section 3. Construction of the PEL ratio confidence intervals involves profiling the pseudo empirical likelihood ratio statistic and is detailed in section 4. All R functions or sample codes are included in the Appendix. They can also be downloaded from the author's personal homepage <http://www.stats.uwaterloo.ca/~cbwu/paper.html>. These functions and codes had been tested in the simulation study reported in Wu and Rao (2004) and were observed to perform very well.

2. Non-Stratified Sampling

Consider a finite population consisting of N identifiable units. Associated with the i^{th} unit are values of the study variable, y_i , and a vector of auxiliary variables, \mathbf{x}_i . The vector of population means $\bar{\mathbf{X}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$ is known. Let $\{(y_i, \mathbf{x}_i), i \in s\}$ be the sample data where s is the set of units selected using a complex survey design. Let $\pi_i = P(i \in s)$ be the inclusion probabilities and $d_i = 1/\pi_i$ be the design weights.

The pseudo empirical maximum likelihood estimator of the population mean $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$ is computed as $\hat{Y}_{\text{PEL}} = \sum_{i \in s} \hat{p}_i y_i$ where the weights \hat{p}_i are obtained by maximizing the pseudo empirical log likelihood function

$$l_{ns}(\mathbf{p}) = n^* \sum_{i \in s} d_i^* \log(p_i) \quad (2.1)$$

subject to the set of constraints

$$0 < p_i < 1, \sum_{i \in s} p_i = 1 \text{ and } \sum_{i \in s} p_i \mathbf{x}_i = \bar{\mathbf{X}}. \quad (2.2)$$

The original pseudo empirical likelihood function proposed by Chen and Sitter (1999) is $l(\mathbf{p}) = \sum_{i \in s} d_i \log(p_i)$. The pseudo empirical likelihood function $l_{ns}(\mathbf{p})$ given by (2.1) was used by Wu and Rao (2004), where $d_i^* = d_i / \sum_{i \in s} d_i$ are the normalized design weights and n^* is the effective sample size. The point estimator $\hat{Y}_{\text{PEL}} = \sum_{i \in s} \hat{p}_i y_i$ remains the same for either version of the likelihood function. The rescaling used in $l_{ns}(\mathbf{p})$ facilitates the construction of the PEL ratio confidence intervals.

Using a standard Lagrange multiplier argument it can be shown that

$$\hat{p}_i = \frac{d_i^*}{1 + \lambda'(\mathbf{x}_i - \bar{\mathbf{X}})} \text{ for } i \in s, \quad (2.3)$$

where the vector-valued Lagrange multiplier, λ , is the solution to

$$g_1(\lambda) = \sum_{i \in s} \frac{d_i^*(\mathbf{x}_i - \bar{\mathbf{X}})}{1 + \lambda'(\mathbf{x}_i - \bar{\mathbf{X}})} = 0.$$

The major computational task here is to find the solution to $g_1(\lambda) = 0$. This can be done using the modified Newton-Raphson procedure proposed by Chen *et al.* (2002). The modification involves checking at each updating stage that the constraint $1 + \lambda'(\mathbf{x}_i - \bar{\mathbf{X}}) > 0$ (i.e., $p_i > 0$) is always satisfied. Without loss of generality, we assume $\bar{\mathbf{X}} = 0$ (if not, replace \mathbf{x}_i by $\mathbf{x}_i - \bar{\mathbf{X}}$ throughout). The modified procedure is as follows.

Step 0: Let $\lambda_0 = \mathbf{0}$. Set $k = 0$, $\gamma_0 = 1$ and $\varepsilon = 10^{-8}$.

Step 1: Calculate $\Delta_1(\lambda_k)$ and $\Delta_2(\lambda_k)$ where

$$\Delta_1(\lambda) = \sum_{i \in s} \frac{d_i^* \mathbf{x}_i}{1 + \lambda' \mathbf{x}_i}$$

and

$$\Delta_2(\lambda) = \left\{ - \sum_{i \in s} d_i^* \frac{\mathbf{x}_i \mathbf{x}_i'}{(\mathbf{x}_i' \lambda)^2} \right\}^{-1} \Delta_1(\lambda).$$

If $\|\Delta_2(\lambda_k)\| < \varepsilon$, stop the algorithm and report λ_k ; otherwise go to Step 2.

Step 2: Calculate $\delta_k = \gamma_k \Delta_2(\lambda_k)$. If $1 + (\lambda_k - \delta_k)' \mathbf{x}_i \leq 0$ for some i , let $\gamma_k = \gamma_k / 2$ and repeat Step 2.

Step 3: Set $\lambda_{k+1} = \lambda_k - \delta_k$, $k = k + 1$ and $\gamma_{k+1} = (\gamma_k + 1)^{-1/2}$. Go to Step 1.

In the original algorithm presented by Chen *et al.* (2002), their step 2 also checks a related dual objective function. While this is necessary for the theoretical proof of convergence of the algorithm, it is not really required for practical applications.

The R function `Lag2(u,ds,mu)` can be used for finding the solution to $g_1(\lambda) = 0$ when the vector of auxiliary variables \mathbf{x} is of dimension m and $m \geq 2$. When \mathbf{x} is univariate, an extremely simple and stable bi-section method to be described shortly should be used. Let n be the sample size. The three required arguments are the $n \times m$ data matrix \mathbf{u} , the $n \times 1$ vector of design weights \mathbf{d}_s and the $m \times 1$ population mean vector μ . The output of the function `Lag2(u,ds,mu)` returns the value of λ which is the solution to $g_1(\lambda) = 0$.

The function `Lag2(u,ds,mu)` will fail to provide a solution if (i) the mean vector $\bar{\mathbf{X}}$ is not an inner point of the convex hull formed by $\{\mathbf{x}_i, i \in s\}$, or (ii) the matrix $\sum_{i \in s} d_i \mathbf{x}_i \mathbf{x}_i'$ is not of full rank. In case (i) the pseudo empirical maximum likelihood estimator does not exist. This happens with probability approaching to zero as the sample size n goes to infinity; in case (ii) one may consider to remove some components of the \mathbf{x} variables from the set of constraints (2.2) to eliminate the collinearity problem.

When the \mathbf{x} variable is univariate, so is the involved Lagrange multiplier λ . In this case we need to solve $g_2(\lambda) = \sum_{i \in s} d_i^* \mathbf{x}_i / (1 + \lambda \mathbf{x}_i) = 0$ for a scalar λ , assuming $\bar{X} = 0$. A unique solution exists if and only if $\min\{x_i, i \in s\} < 0 < \max\{x_i, i \in s\}$. The solution, if exists, lies between $L = -1/\max\{x_i, i \in s\}$ and $U = -1/\min\{x_i, i \in s\}$. Noting that $g_2(\lambda)$ is a monotone decreasing function for $\lambda \in (L, U)$, the most efficient and reliable algorithm for solving $g_2(\lambda) = 0$ is the bi-section method. The function `Lag1(u,ds,mu)` does exactly this, where the required arguments are $\mathbf{u} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{d}_s = (d_1, \dots, d_n)$ and $\mu = \bar{\mathbf{X}}$. The output returns the solution to $g_2(\lambda) = 0$.

The function `Lag1(u,ds,mu)` can be used in conjunction with the model-calibrated pseudo empirical likelihood (MCPEL) approach of Wu and Sitter (2001) to handle cases where the \mathbf{x} variable is high dimensional. The MCPEL approach involves only a single dimension reduction variable derived from a multiple linear regression model and the related Lagrange multiplier problem is always of dimension one.

3. Stratified Sampling

Let $\{(y_{hi}, \mathbf{x}_{hi}), i \in s_h, h = 1, \dots, H\}$ be the sample data from a stratified sampling design. Let $d_{hi}^* = d_{hi} / \sum_{i \in s_h} d_{hi}$ be the normalized design weights for stratum $h, h = 1, \dots, H$. The pseudo empirical likelihood function

under stratified sampling defined by Wu and Rao (2004) is given by

$$l_{st}(\mathbf{p}_1, \dots, \mathbf{p}_H) = n^* \sum_{h=1}^H W_h \sum_{i \in s_h} d_{hi}^* \log(p_{hi}), \quad (3.1)$$

where $W_h = N_h / N$ are the stratum weights and n^* is the total effective sample size as defined in Wu and Rao (2004). The value of n^* is not required for point estimation but this scaling constant is needed for the construction of confidence intervals. Let $\bar{\mathbf{X}}$ be the known vector of population means for auxiliary variables. The maximum pseudo empirical likelihood estimator of the population mean $\bar{\mathbf{Y}} = \sum_{h=1}^H W_h \bar{\mathbf{Y}}_h$ is defined as $\hat{\mathbf{Y}}_{\text{PEL}} = \sum_{h=1}^H W_h \sum_{i \in s_h} \hat{p}_{hi} \mathbf{y}_{hi}$ where the \hat{p}_{hi} maximize $l_{st}(\mathbf{p}_1, \dots, \mathbf{p}_H)$ subject to the set of constraints

$$p_{hi} > 0, \sum_{i \in s_h} p_{hi} = 1, h = 1, \dots, H$$

and

$$\sum_h W_h \sum_{i \in s_h} p_{hi} \mathbf{x}_{hi} = \bar{\mathbf{X}}. \quad (3.2)$$

The major computational difficulty under stratified sampling is caused by the fact that the subnormalization of weights (i.e., $\sum_{i \in s_h} p_{hi} = 1$) occurs at the stratum level while the benchmark constraints (i.e., $\sum_h W_h \sum_{i \in s_h} p_{hi} \mathbf{x}_{hi} = \bar{\mathbf{X}}$) and the constrained maximization of the PEL function are taken at the population level. The algorithm proposed by Wu (2004a) for computing the \hat{p}_{hi} proceeds as follows: let \mathbf{x}_{hi} be augmented to include the first $H-1$ stratum indicator variables and $\bar{\mathbf{X}}$ be augmented to include (W_1, \dots, W_{H-1}) as its first $H-1$ components. In the case of no benchmark constraints involved, the augmented \mathbf{x} variable will consist of the $H-1$ stratum indicator variables only and $\bar{\mathbf{X}} = (W_1, \dots, W_{H-1})$. It follows that the set of constraints (3.2) is equivalent to

$$p_{hi} > 0, \sum_{h=1}^H W_h \sum_{i \in s_h} p_{hi} = 1$$

and

$$\sum_{h=1}^H W_h \sum_{i \in s_h} p_{hi} \mathbf{x}_{hi} = \bar{\mathbf{X}}, \quad (3.3)$$

where the \mathbf{x} variable is now augmented. Let $\mathbf{u}_{hi} = \mathbf{x}_{hi} - \bar{\mathbf{X}}$. It is straightforward by using a standard Lagrange multiplier argument to show that

$$\hat{p}_{hi} = \frac{d_{hi}^*}{1 + \lambda' \mathbf{u}_{hi}},$$

with the vector-valued λ being the solution to

$$g_3(\lambda) = \sum_h W_h \sum_{i \in s_h} \frac{d_{hi}^* \mathbf{u}_{hi}}{1 + \lambda' \mathbf{u}_{hi}} = 0.$$

The modified Newton-Raphson procedure of section 2 for solving $g_1(\lambda) = 0$ can be used for solving $g_3(\lambda) = 0$. The

key computational step under stratified sampling designs is to prepare the data file into suitable format so that the R function `Lag2(u,ds,mu)` for non-stratified sampling can directly be called. Sample R codes for doing this are included in the Appendix.

4. Construction of PEL Ratio Confidence Intervals

While the computational algorithms for the maximum PEL estimator under non-stratified and stratified sampling designs are somewhat different, the search for the lower and the upper boundary of the pseudo empirical likelihood ratio confidence interval for $\bar{\mathbf{Y}}$ involves the same type of profile analysis. Under non-stratified sampling designs, the $(1-\alpha)$ -level PEL ratio confidence interval of $\bar{\mathbf{Y}}$ is constructed as

$$\{\theta \mid r_{ns}(\theta) < \chi_1^2(\alpha)\}, \quad (4.1)$$

where $\chi_1^2(\alpha)$ is the $1-\alpha$ quantile from a χ^2 distribution with one degree of freedom. The pseudo empirical log likelihood ratio statistic $r_{ns}(\theta)$ is computed as

$$r_{ns}(\theta) = -2\{l_{ns}(\tilde{\mathbf{p}}) - l_{ns}(\hat{\mathbf{p}})\},$$

where the $\hat{\mathbf{p}}$ maximize $l_{ns}(\mathbf{p})$ subject to the set of “standard constraints” such as (2.2) and the $\tilde{\mathbf{p}}$ maximize $l_{ns}(\mathbf{p})$ subject to the “standard constraints” plus an additional one induced by the parameter of interest, $\bar{\mathbf{Y}}$, i.e.

$$\sum_{i \in s} p_i y_i = \theta. \quad (4.2)$$

To compute $\tilde{\mathbf{p}}$ one needs to treat (4.2) as an additional component of the “standard constraints” for each fixed value of θ so that the maximization process is essential the same as before.

Let (\hat{L}, \hat{U}) be the interval given by (4.1). Our proposed bi-section method in searching for \hat{L} and \hat{U} is based on following observations:

- i) The minimum value of $r_{ns}(\theta)$ is achieved at $\theta = \sum_{i \in s} \hat{p}_i y_i = \hat{\mathbf{Y}}_{\text{PEL}}$. In this case $\tilde{\mathbf{p}} = \hat{\mathbf{p}}$ and $r_{ns}(\theta) = 0$.
- ii) The interval (\hat{L}, \hat{U}) is bounded by $(y_{(1)}, y_{(n)})$ where $y_{(1)} = \min\{y_i, i \in s\}$ and $y_{(n)} = \max\{y_i, i \in s\}$.
- iii) The pseudo empirical likelihood ratio function $r_{ns}(\theta)$ is monotone decreasing for $\theta \in (y_{(1)}, \hat{\mathbf{Y}}_{\text{PEL}})$ and monotone increasing for $\theta \in (\hat{\mathbf{Y}}_{\text{PEL}}, y_{(n)})$.

Conclusion iii) can be reached by noting that $l_{ns}(\hat{\mathbf{p}})$ does not involve θ and $l_{ns}(\tilde{\mathbf{p}}) = n^* \sum_{i \in s} d_{hi}^* \log(\tilde{p}_i)$ is typically a concave function of θ . It is also possible to show this by directly checking $dr_{ns}(\theta)/d\theta$. For instance, in the case of no auxiliary information involved, the “standard constraints” are $p_i > 0$ and $\sum_{i \in s} p_i = 1$. The \hat{p}_i are given by d_i^* and $\hat{\mathbf{Y}}_{\text{PEL}} = \sum_{i \in s} d_i^* y_i$. The \tilde{p}_i are computed as

$$\tilde{p}_i = \frac{d_i^*}{1 + \lambda(y_i - \theta)}, \quad (4.3)$$

where the λ is the solution to

$$\sum_{i \in s} \frac{d_i^* (y_i - \theta)}{1 + \lambda(y_i - \theta)} = 0. \quad (4.4)$$

Using (4.3) and (4.4), and noting that $\sum_{i \in s} d_i^* / (1 + \lambda(y_i - \theta)) = 1$, it is straightforward to show that

$$\frac{d}{d\theta} r_{ns}(\theta) = 2n^* \sum_{i \in s} \frac{d_i^* \{(d\lambda/d\theta)(y_i - \theta) - \lambda\}}{1 + \lambda(y_i - \theta)} = -2n^* \lambda.$$

By re-writing $d_i^*(y_i - \theta)$ as $d_i^*(y_i - \theta) [(1 + \lambda(y_i - \theta)) - \lambda(y_i - \theta)]$ and after some re-grouping in (4.4) we get

$$\lambda \sum_{i \in s} \frac{d_i^* (y_i - \theta)^2}{1 + \lambda(y_i - \theta)} = \sum_{i \in s} d_i^* y_i - \theta.$$

It follows that $dr_{ns}(\theta)/d\theta = -2n^* \lambda < 0$ if $\theta < \sum_{i \in s} d_i^* y_i = \hat{Y}_{\text{PEL}}$ and $dr_{ns}(\theta)/d\theta > 0$ otherwise.

Sample codes for finding (\hat{L}, \hat{U}) where no auxiliary variable is involved are included in the Appendix. In this case $\hat{p}_i = d_i^*$ and $\hat{Y}_{\text{PEL}} = \sum_{i \in s} d_i^* y_i = \hat{Y}_H$ is the Hajek estimator for \bar{Y} . The profiling process involves finding λ for each chosen value of θ and evaluating the PEL ratio statistic $r_{ns}(\theta)$ against the cut-off value from the χ_1^2 distribution under the desired confidence level $1 - \alpha$. With auxiliary information, one needs to modify the computation of $r_{ns}(\theta)$ for each fixed θ . The bi-section search algorithm for finding \hat{L} and \hat{U} remains the same.

The value of the effective sample size n^* is required for computing the PEL ratio statistic $r_{ns}(\theta)$. For non-stratified sampling designs it is computed as $n^* = \hat{S}_y^2 / \hat{V}(y)$ where

$$\hat{S}_y^2 = \frac{1}{N(N-1)} \sum_{i \in s} \sum_{j > i} \frac{(y_i - y_j)^2}{\pi_{ij}},$$

and

$$\hat{V}(y) = \frac{1}{N^2} \sum_{i \in s} \sum_{j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2,$$

where $e_i = y_i - \hat{Y}_{\text{HT}}$ and $\hat{Y}_{\text{HT}} = N^{-1} \sum_{i \in s} d_i y_i$. See Wu and Rao (2004) for further detail. Computation of n^* involves the second order inclusion probabilities π_{ij} which may impose a real challenge if a π ps sampling scheme is used. In the simulation study reported in Wu and Rao (2004), the Rao-Sampford π ps sampling method was used. R functions for selecting a π ps sample using this method as well as for computing the related second order inclusion probabilities can be found in Wu (2004b). Similar R functions are also available in an add-on R package called “pps”, written by J. Gambino (2003), which can be downloaded from the R homepage (<http://cran.r-project.org/>) by clicking the packages option.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author thanks an associate editor for helpful comments which lead to improvement of the paper.

Appendix: R/S-PLUS Codes

A1. R Function for solving $g_1(\lambda) = 0$.

Let m be the number of auxiliary variables involved and $m \geq 2$. There are three required arguments in the function Lag2(u,ds,mu):

- (1) u: the $n \times m$ data matrix with x_i as its i^{th} row, $i = 1, \dots, n$.
- (2) ds: the $n \times 1$ vector of design weights consisting of d_1, \dots, d_n .
- (3) mu: the $m \times 1$ population mean vector \bar{X} .

The output of the function is the solution to $g_1(\lambda) = 0$.

```
Lag2<-function(u,ds,mu)
{
  n<-length(ds)
  u<-u-rep(1,n)%*%t(mu)
  M<-0*mu
  dif<-1
  tol<-1e-08
  while(dif>tol){
    D1<-0*mu
    DD<-D1%*%t(D1)
    for(i in 1:n){
      aa<-as.numeric(1+t(M)%*%u[i,])
      D1<-D1+ds[i]*u[i,]/aa
      DD<-DD-ds[i]*u[i,]%*%t(u[i,])/aa^2
    }
    D2<-solve(DD,D1,tol=1e-12)
    dif<-max(abs(D2))
    rule<-1
    while(rule>0){
      rule<-0
      if(min(1+t(M-D2)%*%t(u))<=0) rule<-rule+1
      if(rule>0) D2<-D2/2
    }
    M<-M-D2
  }
  return(M)
}
```

A2. R Function for solving $g_2(\lambda) = 0$.

When the x variable is univariate, the solution to $g_2(\lambda) = 0$ can be found through a simple and reliable bi-section method. The three required arguments for the function Lag1(u,ds,mu) are $u = (x_1, \dots, x_n)$, $ds = (d_1, \dots, d_n)$ and $\mu = \bar{X}$. The output is the solution to $g_2(\lambda) = 0$.


```

Lag1<-function(u,ds,mu)
{
  L<--1/max(u-mu)
  R<--1/min(u-mu)
  dif<-1
  tol<-1e-08
  while(dif>tol){
    M<-(L+R)/2
    glam<-sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L<-M
    if(glam<0) R<-M
    dif<-abs(glam)
  }
  return(M)
}

```

A3. Sample code for stratified sampling.

We need to call the function Lag2(u,ds,mu) from nonstratified sampling. The key step is to prepare the data file into suitable format. Let

- (1) $n = (n_1, \dots, n_H)$ be the vector of stratum sample sizes.
- (2) x be the data matrix with x_{hi} as row vectors, $i = 1, \dots, n_h, h = 1, \dots, H$.
- (3) $ds = (d_{11}^*, \dots, d_{1n_1}^*, \dots, d_{H1}^*, \dots, d_{Hn_H}^*)$, where d_{hi}^* are the normalized initial design weights for stratum h .
- (4) X be the vector of known population means.
- (5) $W = (W_1, \dots, W_H)$ be the vector of stratum weights (i.e., $W_h = n_h / N$).

The following sample codes show how the solution to $g_3(\lambda) = 0$ is found (M from the second last line of the following code) and how the \hat{p}_{hi} 's are computed (phi from the last line).

```

###
nst<-sum(n)
k<-length(n)-1
ntot<-rep(0,k)
ntot[1]<-n[1]
for(j in 2:k) ntot[j]<-ntot[j-1]+n[j]
ist<-matrix(0,nst,k)
ist[,1,n[1],1]<-1
for(j in 2:k) ist[(ntot[j-1]+1):ntot[j],j]<-1
uhi<-cbind(ist,x)
mu<-c(W[1,k],X)
whi<-rep(W[1],n[1])
for(j in 2:(k+1)) whi<-c(whi,rep(W[j],n[j]))
dhi<-whi*ds
M<-Lag2(uhi,dhi,mu)
phi<-as.vector(ds/(1+(uhi-rep(1,nst)%*%(mu))%*%M))
###

```

A4. Sample code for finding the PEL ratio confidence interval.

The search for the lower boundary (LB) and the upper boundary (UB) of the PEL ratio confidence interval needs to be carried out separately. The following codes show how this is done for the case of no auxiliary information. With auxiliary information, one needs to modify the computation

of the involved pseudo empirical likelihood ratio statistic (elratio) accordingly. Let

- (1) $\alpha = 1 - \alpha$ be the confidence level of the desired interval.
- (2) $ys = (y_1, \dots, y_n)$ be the sample data.
- (3) $ds = (d_1^*, \dots, d_n^*)$ be the normalized design weights.
- (4) $YEL = \sum_{i \in s} \hat{p}_i y_i$ (in this case $\hat{p}_i = d_i^*$).
- (5) nss be the estimated effective sample size n^* .

```

###
tol<-1e-08
cut<-qchisq(a,1)
###
t1<-YEL
t2<-max(ys)
dif<-t2-t1
while(dif>tol){
  tau<-(t1+t2)/2
  M<-Lag1(ys,ds,tau)
  elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
  if(elratio>cut) t2<-tau
  if(elratio<=cut) t1<-tau
  dif<-t2-t1
}
UB<-(t1+t2)/2
###
t1<-YEL
t2<-min(ys)
dif<-t1-t2
while(dif>tol){
  tau<-(t1+t2)/2
  M<-Lag1(ys,ds,tau)
  elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
  if(elratio>cut) t2<-tau
  if(elratio<=cut) t1<-tau
  dif<-t1-t2
}
LB<-(t1+t2)/2
###

```

References

- Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.
- Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- Wu, C. (2004a). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.
- Wu, C. (2004b). R/S-PLUS Implementation of pseudo empirical likelihood methods under unequal probability sampling. Working paper 2004-07, Department of Statistics and Actuarial Science, University of Waterloo.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- Wu, C., and Rao, J.N.K. (2004). Pseudo empirical likelihood ratio confidence intervals for complex surveys. Working paper 2004-06, Department of Statistics and Actuarial Science, University of Waterloo.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2005.

- | | |
|---|--|
| A.K. Adhikary, <i>ISI Kolkata</i> | B. Mandall, <i>Ohio State University</i> |
| M. Battaglia, <i>ABT Associates</i> | S. Matthews, <i>Statistics Canada</i> |
| J.-F. Beaumont, <i>Statistics Canada</i> | D. Marker, <i>Westat, Inc.</i> |
| N. Billor, <i>Auburn University</i> | D. McCaffrey, <i>RAND</i> |
| A. Boudreau, <i>Medical College of Wisconsin</i> | C.E. M'LAN, <i>University of Connecticut</i> |
| K. Brewer, <i>Australian National University</i> | J. Moore, <i>U.S. Bureau of the Census</i> |
| F. Butar Butar, <i>Sam Houston State University</i> | R. Munnich, <i>University of Tübingen</i> |
| D. Cantor, <i>Westat</i> | J. Opsomer, <i>Iowa State University</i> |
| S.R. Chowdhury, <i>Westat, Inc.</i> | O. Phillips, <i>Statistics Canada</i> |
| S.L. Christ, <i>University of North Carolina</i> | M. Pratesi, <i>University of Pisa, Italy</i> |
| J. Chromy, <i>RTI International</i> | J. Reiter, <i>Duke University</i> |
| R. Courtemanche, <i>Institut de la statistique du Québec</i> | R.H. Renssen, <i>Statistics Netherlands</i> |
| A. Dessertaine, <i>EDF R&D-OSIRIS - CLAMART</i> | G. Roberts, <i>Statistics Canada</i> |
| P. Dick, <i>Statistics Canada</i> | I. Şchiopu-Kratina, <i>Statistics Canada</i> |
| P. Duchesne, <i>Université de Montréal</i> | C.J. Schwarz, <i>Simon Fraser University</i> |
| J. Dumais, <i>Statistics Canada</i> | A. Scott, <i>University of Auckland</i> |
| J. Eltinge, <i>Bureau of Labour Statistics</i> | J. Sedransk, <i>Case Western University</i> |
| M. Feder, <i>Research Triangle Institute</i> | R. Sitter, <i>Simon Fraser University</i> |
| R. Fisher, <i>U.S. Census Bureau</i> | M. Sinclair, <i>U.S. Department of Labor</i> |
| O. Frank, <i>Stockholm University</i> | A. Singh, <i>Statistics Canada</i> |
| S.G. Heeringa, <i>Institute for Social Research, University of Michigan</i> | T.W. Smith, <i>NORC</i> |
| S. Haslett, <i>Massey University</i> | J. Stec, <i>InteCap, Inc.</i> |
| D. Heng-Yan Leung, <i>Singapore Management University</i> | D.G. Steel, <i>University of Wollongong, Australia</i> |
| K. Jae Kwang, <i>Yonsei University</i> | L. Stokes, <i>Southern Methodist University</i> |
| F. Jenkins, <i>Westat</i> | E. Stuart, <i>Mathematica Policy Research, Inc.</i> |
| J. Jiang, <i>University of California at Davis</i> | A. Jr. Tersine, <i>United States Bureau of the Census</i> |
| J.K. Kim, <i>Yonsei University</i> | R. Thomas, <i>Carleton University</i> |
| M. Kovačević, <i>Statistics Canada</i> | N. Thomas, <i>Pfizer, Inc.</i> |
| S. Laaksonen, <i>University of Helsinki and Statistics Finland</i> | C. Tucker, <i>United States Bureau of Labor</i> |
| P. Lahiri, <i>University of Maryland</i> | J. van der Brakel, <i>Statistics Netherlands</i> |
| F. Lapointe, <i>Institut de la statistique du Québec</i> | S.L. Vartivarian, <i>Mathematica Policy Research, Inc.</i> |
| M.D. Larsen, <i>Iowa State University</i> | J. Wang, <i>Merck Research Labs, Merck & Co., Inc.</i> |
| P. Lavallée, <i>Statistics Canada</i> | X. Wang, <i>Southern Methodist University</i> |
| H. Lee, <i>Westat, Inc.</i> | C. Wu, <i>University of Waterloo</i> |
| R. Lehtonen, <i>University of Jyväskylä</i> | R. Yucel, <i>University of Massachusetts</i> |
| N.T. Longford, <i>SNTL</i> | W. Yung, <i>Statistics Canada</i> |
| L. Magee, <i>McMaster University</i> | E. Zanutto, <i>University of Pennsylvania</i> |
| T. Maiti, <i>Iowa State University</i> | H. Zheng, <i>Massachusetts General Hospital and Harvard Medical School</i> |
| D. Malec, <i>United States Bureau of the Census</i> | |

Acknowledgements are also due to those who assisted during the production of the 2005 issues: Francine Pilon-Renaud and Roberto Guido (Dissemination Division), Marc Bazinet (Marketing Division) and François Beaudin (Official Languages and Translation Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier, Nancy Flansberry and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 21, No. 2, 2005

Reflections on Early History of Official Statistics and a Modest Proposal for Global Coordination Samuel Kotz	139
The Effectiveness of a Supranational Statistical Office Pluses, Minuses, and Challenges Viewed from the Outside Ivan P. Fellegi and Jacob Ryten.....	145
An Interview with the Authors of the Book <i>Model Assisted Survey Sampling</i> Phillip S. Kott, Bengt Swensson, Carl-Erik Särndal, and Jan Wretman	171
Achieving Usability in Establishment Surveys Through the Application of Visual Design Principles Don A. Dillman, Arina Gertseva, and Taj Mahon-Haft.....	183
Promoting Uniform Question Understanding in Today's and Tomorrow's Surveys Frederick G. Conrad and Michael F. Schober	215
To Mix or Not to Mix Data Collection Modes in Surveys Edith deLeeuw	233
Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project Jeroen Pannekoek and Ton de Waal.....	255
Model-based Estimation of Drug Use Prevalence Using Item Count Data Paul P. Biemer and Gordon Brown.....	285
Data Swapping: Variations on a Theme by Dalenius and Reiss Stephen E. Fienberg and Julie McIntyre	307
PRIMA: A New Multiple Imputation Procedure for Binary Variables Ralf Münnich and Susanne Rässler.....	323
Some Recent Developments and Directions in Seasonal Adjustment David F. Findley	341

Volume 21, No. 3, 2005

Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects Robert J.J. Voogt and Willem E. Saris.....	367
Separating Interviewer and Sampling-Point Effects Rainer Schnell and Frauke Kreuter	389
Small Area Estimation from the American Community Survey Using a Hierarchical Logistic Model of Persons and Housing Units Donald Malec.....	411
A Note on the Hartley-Rao Variance Estimator Phillip S. Kott.....	433
Using CART to Generate Partially Synthetic Public Use Microdata J.P. Reiter	441
Purchasing Power Parity Measurement and Bias from Loose Item Specifications in Matched Samples: An Analytical Model and Empirical Study Mick Silver and Saeed Heravi	463
Estimating the Number of Distinct Valid Signatures in Initiative Petitions Ruben A. Smith and David R. Thomas.....	489
Official Statistics in Hungary Before Full Membership in the EU Tamas Mellár	505
Book and Software Reviews.....	517
In Other Journals.....	527

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 33, No. 2, June/juin 2005

David HAZIZA & J.N.K. RAO	
Inference for domains under imputation for missing survey data	149
Camelia GOGA	
Variance reduction in surveys with auxiliary information: a nonparametric approach involving regression splines	163
María-José LOMBARDÍA, Wenceslao GONZÁLEZ-MANTEIGA & José-Manuel PRADA-SÁNCHEZ	
Estimation of a finite population distribution function based on a linear model with unknown heteroscedastic errors	181
Todd MACKENZIE & Michal ABRAHAMOWICZ	
Using categorical markers as auxiliary variables in log-rank tests and hazard ratio estimation	201
Holger DETTE, Linda M. HAINES & Lorens A. IMHOF	
Bayesian and maximin optimal designs for heteroscedastic regression models	221
Jennifer ASIMIT & W. John BRAUN	
Third order point process intensity estimation for reaction time experiment data	243
W. John BRAUN & Li-Shan HUANG	
Kernel spline regression	259
Mario FRANCISCO-FERNANDEZ & Jean D. OPSOMER	
Smoothing parameter selection methods for nonparametric regression with spatially correlated errors	279
Zeny Z. FENG, Jiahua CHEN & Mary E. THOMPSON	
The universal validity of the possible triangle constraint for affected sib pairs	297
Forthcoming papers/Articles à paraître	311

Volume 33, No. 3, September/septembre 2005

Preface/Préface	313
Belkacem ABDOUS, Anne-Laure FOUGÈRES & Kilani GHOUDI	
Extreme behaviour for bivariate elliptical distributions	317
Yinshan ZHAO & Harry JOE	
Composite likelihood estimation in multivariate data analysis	335
Hideatsu TSUKAHARA	
Semiparametric estimation in copula models	357
François VANDENHENDE & Philippe LAMBERT	
Local dependence estimation using semiparametric Archimedean copulas	377
Xiaohong CHEN & Yanqin FAN	
Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection	389
Olivier SCAILLET	
A Kolmogorov-Smirnov type test for positive quadrant dependence	415
Roel BRAEKERS & Noël VERAVERBEKE	
A copula-graphic estimator for the conditional survival function under dependent	429
Yun-Hee CHOI & David E. MATTHEWS	
Accelerated life regression modelling of dependent bivariate time-to-event data	449
David OAKES	
On the preservation of copula structure under truncation	465

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, N° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
- 1.1

Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2

Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3

Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4

Les remerciements doivent paraître à la fin du texte.
- 1.5

Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. **Résumé**
- Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
- 3.1

Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2

Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3

Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4

Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5

Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; 1).
- 3.6

Les caractères italiques sont utilisés pour faire ressortir des mots.

4. **Figures et tableaux**
- 4.1

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2

Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. **Bibliographie**
- 5.1

Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
- 5.2

La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Volume 33, No. 2, June/juin 2005

David HAZIZA & J.N.K. RAO	Inference for domains under imputation for missing survey data	149
Camelia GOGA	Variance reduction in surveys with auxiliary information: a nonparametric approach involving regression splines.....	163
Marta-José LOMBARDA, Wenceslao GONZÁLEZ-MANTEIGA & José-Manuel PRADA-SANCHEZ	Estimation of a finite population distribution function based on a linear model with unknown heteroscedastic errors	181
Todd MACKENZIE & Michal ABRAMOWICZ	Using categorical markers as auxiliary variables in log-rank tests and hazard ratio estimation.....	201
Holger DETTE, Linda M. HAINES & Lorens A. IMHOF	Bayesian and maximin optimal designs for heteroscedastic regression models.....	221
Jennifer ASMIT & W. John BRAUN	Third order point process intensity estimation for reaction time experiment data.....	243
W. John BRAUN & Li-Shan HUANG	Kernel spline regression.....	259
Matéo FRANCISCO-FERNANDEZ & Jean D. OPSOMER	Smoothing parameter selection methods for nonparametric regression with spatially correlated errors.....	279
Zeny Z. FENG, Jiahua CHEN & Mary E. THOMPSON	The universal validity of the possible triangle constraint for affected sib pairs	297
	Forthcoming papers/Articles à paraître	311

Volume 33, No. 3, September/septembre 2005

	Preface/Préface.....	313
Belkacem ABDOUN, Anne-Laure FOUGERES & Kilani GHOUDI	Extreme behaviour for bivariate elliptical distributions.....	317
Yinshan ZHAO & Harry JOE	Composite likelihood estimation in multivariate data analysis	335
Hidetatsu TSUKAHARA	Semiparametric estimation in copula models.....	357
François VANDENHENDEN & Philippe LAMBERT	Local dependence estimation using semiparametric Archimedean copulas.....	377
Xiaohong CHEN & Yanqin FAN	Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection.....	389
Olivier SCALLET	A Kolmogorov-Smirnov type test for positive quadrant dependence.....	415
Roel BRAEKERS & Noël VERAVERBEKE	A copula-graphic estimator for the conditional survival function under dependent.....	429
Yun-Hee CHOI & David E. MATTHEWS	Accelerated life regression modelling of dependent bivariate time-to-event data.....	449
David OAKES	On the preservation of copula structure under truncation.....	465

Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects	Robert J. J. Voogt and Willem E. Saris	367
Separating Interviewer and Sampling-Point Effects	Rainer Schnell and Frauke Kreuter	389
Small Area Estimation from the American Community Survey Using a Hierarchical Logistic Model of Persons and Housing Units	Donald Malec	411
A Note on the Hartley-Rao Variance Estimator	Phillip S. Kott	433
Using CART to Generate Partially Synthetic Public Use Microdata	J. P. Reiter	441
Purchasing Power Parity Measurement and Bias from Loose Item Specifications in Matched Samples:	An Analytical Model and Empirical Study	
	Mick Silver and Saeed Heravi	463
Estimating the Number of Distinct Valid Signatures in Initiative Petitions	Ruben A. Smith and David R. Thomas	489
Official Statistics in Hungary Before Full Membership in the EU	Tamas Mellár	505
Book and Software Reviews		517
In Other Journals		527

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

Contents Volume 21, No. 2, 2005

Reflections on Early History of Official Statistics and a Modest Proposal for Global Coordination	Samuel Kotz	139
The Effectiveness of a Supranational Statistical Office Pluses, Minuses, and Challenges Viewed from the Outside	Ivan P. Fellegi and Jacob Rytén.....	145
An Interview with the Authors of the Book <i>Model Assisted Survey Sampling</i>	Phillip S. Kott, Bengt Swensson, Carl-Erik Särndal, and Jan Wretman.....	171
Achieving Usability in Establishment Surveys Through the Application of Visual Design Principles	Don A. Dillman, Arina Gertseva, and Tay Mahon-Hafl.....	183
Promoting Uniform Question Understanding in Today's and Tomorrow's Surveys	Frederick G. Conrad and Michael F. Schober.....	215
To Mix or Not to Mix Data Collection Modes in Surveys	Edith de Leeuw	233
Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project	Jeroen Pannekoek and Ton de Waal.....	255
Model-based Estimation of Drug Use Prevalence Using Item Count Data	Paul P. Biemer and Gordon Brown.....	285
Data Swapping: Variations on a Theme by Dalenius and Reiss	Stephen E. Fienberg and Julie McInyre.....	307
PRIMA: A New Multiple Imputation Procedure for Binary Variables	Ralf Munnich and Susanne Rässler.....	323
Some Recent Developments and Directions in Seasonal Adjustment	David F. Findley.....	341

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique ou plus durant l'année 2005.

A.K. Adhikary, ISI Kolkata
 M. Battaglia, ABT Associates
 J.-F. Beaumont, Statistique Canada
 N. Billor, Auburn University
 C. Boudreau, Medical College of Wisconsin
 K. Brewer, Australian National University
 F. Butar Butar, Sam Houston State University
 D. Cantor, Westat
 S.R. Chowdhury, Westat, Inc.
 S.L. Christ, University of North Carolina
 J. Chromy, RTI International
 R. Courtemanche, Institut de la statistique du Québec
 A. Dessertaine, EDF R&D-OSIRIS - CLAMART
 P. Dick, Statistique Canada
 P. Duchesne, Université de Montréal
 J. Dumais, Statistique Canada
 J. Eltinge, Bureau of Labour Statistics
 M. Feder, Research Triangle Institute
 R. Fisher, U.S. Census Bureau
 O. Frank, Stockholm University
 S.G. Heeringa, Institute for Social Research, University of Michigan
 S. Haslett, Massey University
 D. Heng-Yan Leung, Singapore Management University
 K. Jae Kwang, Yonsei University
 F. Jenkins, Westat
 J. Jiang, University of California at Davis
 J.K. Kim, Yonsei University
 M. Kovacevic, Statistique Canada
 S. Laaksonen, University of Helsinki and Statistics Finland
 P. Lahiri, University of Maryland
 F. Lapointe, Institut de la statistique du Québec
 M.D. Larsen, Iowa State University
 P. Lavallée, Statistique Canada
 H. Lee, Westat, Inc.
 R. Lehtonen, University of Jyväskylä
 N.T. Longford, SNTL
 L. Magee, McMaster University
 T. Maiti, Iowa State University
 D. Malec, United States Bureau of the Census

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2005: Franchné Pilon-Renaud et Roberto Guido (Division de la diffusion), Marc Bazinet (Division du marketing) et François Beaudin (Division des langues officielles et traduction). Finalement nous désirons exprimer notre reconnaissance à Christine Cousineau, Céline Ethier, Nancy Flansberry et Denis Lemire de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

B. Mandall, Ohio State University
 S. Mathews, Statistique Canada
 D. Marker, Westat, Inc.
 D. McCaffrey, RAND
 C.E. M'Lan, University of Connecticut
 J. Moore, U.S. Bureau of the Census
 R. Munnich, University of Tübingen
 J. Opsomer, Iowa State University
 O. Phillips, Statistique Canada
 M. Pratesi, University of Pisa, Italy
 J. Reiter, Duke University
 R.H. Renssen, Statistisches Neherlands
 G. Roberts, Statistique Canada
 I. Schiopu-Kratina, Statistique Canada
 C.J. Schwarz, Simon Fraser University
 A. Scott, University of Auckland
 J. Sedransk, Case Western University
 R. Sitter, Simon Fraser University
 M. Sinclair, U.S. Department of Labor
 A. Singh, Statistique Canada
 T.W. Smith, NORC
 J. Siec, InteCap, Inc.
 D.G. Steel, University of Wollongong, Australia
 L. Stokes, Southern Methodist University
 E. Stuart, Mathematica Policy Research, Inc.
 A. Jr. Terstine, United States Bureau of the Census
 R. Thomas, Carleton University
 N. Thomas, Pfizer, Inc.
 C. Tucker, United States Bureau of Labor
 J. van der Brakel, Statistisches Neherlands
 S.L. Vartivarian, Mathematica Policy Research, Inc.
 I. Wang, Merck Research Labs, Merck & Co., Inc.
 C. Wu, University of Waterloo
 R. Yucel, University of Massachusetts
 W. Yung, Statistique Canada
 E. Zanutto, University of Pennsylvania
 H. Zheng, Massachusetts General Hospital and Harvard Medical School

Bibliographie

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Wu, C. (2004a). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2004b). R/S-PLUS Implementation of pseudo empirical likelihood methods under unequal probability sampling. Document de travail 2004-07, Department of Statistics and Actuarial Science, University of Waterloo.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the Americal Statistical Association*, 96, 185-193.

Wu, C., et Rao, J.N.K. (2004). Pseudo empirical likelihood ratio confidence intervals for complex surveys. Document de travail 2004-06, Department of Statistics and Actuarial Science, University of Waterloo.

```
###
toI<-1e-08
cut<-q-phisq(a,1)
###
l1<-YEL
l2<-max(ys)
diff<-l2-l1
while(diff>toI){
  tau<-(t1+l2)/2
  M<-Lag1(ys,ds,tau)
  erratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
  if(erratio>cut) l2<-tau
  diff<-l2-l1
}
UB<-(t1+l2)/2
###
l1<-YEL
l2<-min(ys)
diff<-l1-l2
while(diff>toI){
  tau<-(t1+l2)/2
  M<-Lag1(ys,ds,tau)
  erratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
  if(erratio>cut) l2<-tau
  diff<-l2-l1
}
LB<-(t1+l2)/2
###
```

Annexe : Codes RS-PLUS

A1. Fonction R pour résoudre $g_1(\lambda) = 0$.

Soit m le nombre de variables auxiliaires concernés et $m \geq 2$. Trois arguments sont requis dans la fonction $\text{Lag2}(u, ds, mu)$:

- (1) u : la matrice de données de dimensions $n \times m$ avec x_i en tant que i^{e} ligne $i = 1, \dots, n$;
- (2) ds : le vecteur de poids de sondage de dimension $n \times 1$ constitué de d_1, \dots, d_n ;
- (3) mu : le vecteur de moyennes de population de dimension $m \times 1$ \bar{X} .

La sortie de la fonction est la solution de $g_1(\lambda) = 0$.

```
Lag2<-function(u,ds,mu)
{
  n=length(ds)
  M<-u-rep(1,n)%*%(mu)
  M<-0*mu
  tolt<-1e-08
  while(diff>tol)
  {
    D1<-0*mu
    DD<-D1*%*%(D1)
    for(i in 1:n)
    {
      aac<-as.numeric(1+(M*mu[i]))
      D1<-D1+ds[i]*u[i,]/aac
      DD<-DD+ds[i]*u[i,]%*%t(u[i,])/aac2
    }
    D2<-solve(DD,D1,lolet=1e-12)
    diff<-max(abs(D2))
    rule<-1
    while(rule>0)
    {
      rule<-0
      if(m1(1+(M-D2*%*%(u))<=0) rule<-rule+1
      if(rule>0) D2<-D2/2
    }
    M<-M-D2
  }
  return(M)
}
```

A2. Fonction R pour la résolution de $g_2(\lambda) = 0$.

Lorsque la variable x est univariée, la solution de $g_2(\lambda) = 0$ peut être obtenue au moyen d'une méthode de bisection simple et fiable. Les trois arguments requis pour la fonction $\text{Lag1}(u, ds, mu)$ sont $u = (x_1, \dots, x_n)$, $ds = (d_1, \dots, d_n)$ et $mu = \bar{X}$. La sortie est la solution de $g_2(\lambda) = 0$.

```
Lag1<-function(u,ds,mu)
{
  L<-1/max(u-mu)
  R<-1/min(u-mu)
  tolt<-1e-08
  while(diff>tol)
  {
    M<-(L+R)/2
    glam<-sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L<-M
    if(glam<0) R<-M
    diff<-abs(glam)
  }
  return(M)
}
```

Les exemples de codes qui suivent montrent comment est trouvée la solution de $g_3(\lambda) = 0$ (M de l'avant-dernière ligne du code qui suit) et comment sont calculés les \hat{p}_{hi} (phi de la dernière ligne).

(c'est-à-dire $W^h = N^h / N$).

- (4) X le vecteur de moyennes de population connues ;
- (5) $W = (W_1, \dots, W_H)$ le vecteur de poids de strate pour la strate h ;
- (3) $ds = (d_{11}^*, \dots, d_{1n_1}^*, \dots, d_{H1}^*, \dots, d_{Hn_H}^*)$, où les d_{hi}^* sont les poids de sondage initiaux normalisés de l'ligne, $i = 1, \dots, n_h$, $h = 1, \dots, H$;
- (2) x la matrice de données avec x_{hi} comme vecteurs de strate ;
- (1) $n = (n_1, \dots, n_H)$ le vecteur de taille d'échantillon

A3. Exemple de code pour l'échantillonnage stratifié. Nous devons appeler la fonction $\text{Lag2}(u, ds, mu)$ à partir de l'échantillonnage non stratifié. L'étape essentielle est la préparation du fichier de données afin de lui donner le format approprié. Soit

A4. Exemple de code pour trouver l'intervalle de confiance du rapport de pseudo-vraisemblance empirique. La recherche de la borne inférieure (LB) et de la borne supérieure (UB) de l'intervalle de confiance du rapport de vraisemblance empirique doit se faire séparément. Les codes qui suivent montrent comment se fait cette recherche dans le cas où l'on n'utilise aucune information auxiliaire. Si l'on utilise ce genre d'information, il faut modifier le calcul des rapports de pseudo-vraisemblance empirique concernés (c'est-à-dire) en conséquence. Soit

- (1) $a = 1 - \alpha$ le niveau de confiance de l'intervalle souhaité ;
- (2) $ys = (y_1^n, \dots, y_n^n)$ les données d'échantillon ;
- (3) $ds = (d_1^*, \dots, d_n^*)$ les poids de sondage normalisés ;
- (4) $YEL = \sum_{i \in s} \hat{p}_i y_i$ (ici $\hat{p}_i = d_i^*$) ;
- (5) nss la taille d'échantillon effective estimée n^* .

où $\chi^2_2(\alpha)$ est le quantile $1 - \alpha$ d'une loi χ^2_2 à un degré de liberté. Le rapport des log pseudo-vraisemblances empiriques $r_{ns}(\theta)$ est donné par

$$r_{ns}(\theta) = -2[l_{ns}(\bar{p}) - l_{ns}(\bar{p})],$$

où les \bar{p} maximisent $l_{ns}(\bar{p})$ sous l'ensemble de « contraintes standard » telles que (2.2) et les \bar{p} maximisent $l_{ns}(\bar{p})$ sous les « contraintes standard » et une contrainte supplémentaire induite par le paramètre d'intérêt, Y_i , c'est-à-dire

$$\sum_{i \in s} p_i Y_i = \theta. \quad (4.2)$$

Pour calculer \bar{p} , il faut traiter (4.2) comme une composante supplémentaire de l'ensemble de « contraintes standard » pour chaque valeur fixée de θ , de sorte que le processus de maximisation soit essentiellement le même qu'auparavant.

Soit (\bar{L}, \bar{U}) l'intervalle donné par (4.1). La méthode de bisection que nous avons proposée pour trouver \bar{L} et \bar{U} est fondée sur les observations suivantes :

- i) La valeur minimale de $r_{ns}(\theta)$ est atteinte à $\theta = \sum_{i \in s} \bar{p}_i Y_i = Y_{\text{PEL}}$. Dans ce cas, $\bar{p} = \bar{p}$ et $r_{ns}(\theta) = 0$.
- ii) L'intervalle (\bar{L}, \bar{U}) est borné par $(Y^{(1)}, Y^{(n)})$, où $Y^{(1)} = \min\{Y_i, i \in s\}$ et $Y^{(n)} = \max\{Y_i, i \in s\}$.
- iii) Le rapport de pseudo-vraisemblance empirique $r_{ns}(\theta)$ est une fonction monotone décroissante pour $\theta \in (\bar{Y}^{(1)}, \bar{Y}^{(\text{PEL})})$ et monotone croissante pour $\theta \in (\bar{Y}^{(\text{PEL})}, Y^{(n)})$.

Nous pouvons arriver à la conclusion iii) en notant que $l_{ns}(\bar{p})$ ne fait pas intervenir θ et que $l_{ns}(\bar{p}) = \sum_{i \in s} \bar{p}_i \log(\bar{p}_i)$ est typiquement une fonction concave de θ . Il est également possible de montrer cela en vérifiant directement $dr_{ns}(\theta)/d\theta$. Par exemple, dans le cas où n'intervient aucune information auxiliaire, les « contraintes standard » sont $p_i > 0$ et $\sum_{i \in s} p_i = 1$. Les \bar{p}_i sont donnés par d_i^* et $Y_{\text{PEL}} = \sum_{i \in s} d_i^* Y_i$. Les \bar{p}_i sont calculés sous la forme

$$\bar{p}_i = \frac{d_i^*}{d_i^* + 1 + \lambda(Y_i - \theta)}, \quad (4.3)$$

où λ est la solution de

$$\sum_{i \in s} \frac{d_i^* (Y_i - \theta)}{1 + \lambda(Y_i - \theta)} = 0. \quad (4.4)$$

En partant de (4.3) et (4.4), et en notant que $\sum_{i \in s} d_i^* / (1 + \lambda(Y_i - \theta)) = 1$, il est facile de montrer que

$$\frac{d}{d\theta} r_{ns}(\theta) = 2n^* \sum_{i \in s} d_i^* \frac{\{d_i^* \lambda / (d\theta)(Y_i - \theta) - \lambda\}}{1 + \lambda(Y_i - \theta)} = -2n^* \lambda.$$

En réécrivant $d_i^* (Y_i - \theta)$ sous la forme $d_i^* (Y_i - \theta) [1 + \lambda(Y_i - \theta)] - \lambda(Y_i - \theta)$ et après certains regroupements dans (4.4), nous obtenons

Remerciements

Cette étude a été financée par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. L'auteur remercie un rédacteur associé de ses commentaires constructifs qui lui ont permis d'améliorer l'article.

en cliquant sur l'option *packages*.

et $e_i = Y_i - Y_{\text{HT}} - Y_{\text{HT}} = N^{-1} \sum_{i \in s} d_i^* Y_i$. Consulter Wu (2004) pour plus de précisions. Le calcul de n^* comprend les probabilités de sélection de deuxième ordre π_{ij} qui peuvent poser un vrai défi si l'on utilise un plan de sondage PPT. Dans leur étude en simulation, Wu et Rao (2004) ont utilisé la méthode d'échantillonnage PPT de Rao-Sampford. Les fonctions R pour sélectionner un échantillon PPT selon cette méthode, ainsi que pour calculer les probabilités de sélection de deuxième ordre connexes peuvent être consultées dans Wu (2004b). Des fonctions R similaires sont également disponibles dans un projet R complémentaire appelé « pps » [pour *probability proportional to size*], rédigé par J. Gambino (2003), qui peut être téléchargé à la page d'accueil R à <http://cran.r-project.org/>

$$V(Y) = \frac{1}{N} \sum_{i \in s} \sum_{j > i} \pi_{ij} \left(\frac{\pi_{ij}}{e_i} - \frac{\pi_{ij}}{e_j} \right)^2,$$

$$S_y^2 = \frac{1}{N(N-1)} \sum_{i \in s} \sum_{j > i} (Y_i - Y_j)^2 \pi_{ij},$$

la calcule selon $n^* = S_y^2 / V(Y)$, où $r_{ns}(\theta)$. Dans la cas des plans de sondage non stratifiés, on calcule le rapport de pseudo-vraisemblance empirique pour La taille effective d'échantillon n^* doit être connue pour l'algorithm de bisection pour trouver \bar{L} et \bar{U} demeure le même.

modifier le calcul de $r_{ns}(\theta)$ pour chaque valeur fixée de θ . souhaité. Si l'on utilise des données auxiliaires, il faut seuil de la loi χ^2_2 sous le niveau de confiance $1 - \alpha$ vraisemblance empirique $r_{ns}(\theta)$ en fonction de la valeur valeur choisie de θ et à évaluer le rapport de pseudo-bissemment du profil consiste à trouver λ pour chaque blissement l'estimateur de Hajek de Y . Le processus d'éta-Dans ces conditions, $\bar{p}_i = d_i^*$ et $Y_{\text{PEL}} = \sum_{i \in s} d_i^* Y_i = Y_H$ aucune variable auxiliaire n'est utilisée figurent à l'annexe. Les exemples de codes pour trouver (\bar{L}, \bar{U}) quand $\sum_{i \in s} d_i^* Y_i = Y_{\text{PEL}}$ et $dr_{ns}(\theta)/d\theta > 0$ autrement.

Il s'ensuit que $dr_{ns}(\theta)/d\theta = -2n^* \lambda < 0$ si $\theta < \sum_{i \in s} d_i^* Y_i = Y_{\text{PEL}}$ et $dr_{ns}(\theta)/d\theta = -2n^* \lambda > 0$ si $\theta > \sum_{i \in s} d_i^* Y_i = Y_{\text{PEL}}$.

$$\lambda = \frac{\sum_{i \in s} d_i^* (Y_i - \theta)}{1 + \lambda(Y_i - \theta)} = \sum_{i \in s} d_i^* Y_i - \theta.$$

$\sum_h W_h \sum_{i \in s_h} p_{hi} x_{hi} = \bar{X}$) et la maximisation contrainte au niveau de la population. L'algorithme proposé par Wu (2004a) pour calculer les \hat{p}_{hi} se déroule comme suit : soit l'augmentation de x_{hi} afin d'inclure les $H-1$ premières variables indicatrices de strate et l'augmentation de \bar{X} afin d'inclure (W_1, \dots, W^{H-1}) en tant que ses $H-1$ premières composantes. Dans le cas où il n'existe aucune contrainte d'échantillonnage, la variable x augmentée correspond aux $H-1$ variables indicatrices de strate uniquement et $\bar{X} = (W_1, \dots, W^{H-1})$. Il s'ensuit que l'ensemble de contraintes (3.2) est équivalent à

$$p_{hi} > 0, \sum_{h=1}^H W_h \sum_{i \in s_h} p_{hi} = 1$$

(3.3)

où la variable x est maintenant augmentée. Soit $u_{hi} = x_{hi} - \bar{X}$. Il est facile, en utilisant comme argument un multiplicateur de Lagrange standard, de montrer que

$$\hat{p}_{hi} = \frac{d_{hi}^* \chi_{hi}}{1 + d_{hi}^* \chi_{hi}},$$

où χ_{hi} évalué vectoriellement est la solution de

$$g_3(\chi) = \sum_h W_h \sum_{i \in s_h} \frac{d_{hi}^* \chi_{hi}}{1 + d_{hi}^* \chi_{hi}} = 0.$$

La procédure de Newton-Raphson modifiée de la section 2 pour la résolution de $g_1(\chi) = 0$ peut être utilisée sous résoudre $g_3(\chi) = 0$. L'étape clé du calcul sous l'échantillonnage stratifié consiste à donner au fichier de données un format approprié pour pouvoir appeler directement la fonction `R Lag2(u,d,s,mu)` utilisée pour l'échantillonnage non stratifié. Des exemples de codes R pour le faire figurent en annexe.

4. Construction des intervalles de confiance des rapports de pseudo-vraisemblance empirique

Bien que les algorithmes informatiques pour l'estimateur du maximum de vraisemblance pseudo empirique sous plans de sondage stratifié et non stratifié diffèrent quelque peu, la recherche des bornes inférieure et supérieure de l'intervalle de confiance du rapport de pseudo-vraisemblance empirique pour \bar{Y} comporte le même type d'analyse de profil. Sous un plan de sondage non stratifié, l'intervalle de confiance de niveau $(1 - \alpha)$ du rapport de pseudo-vraisemblance empirique de \bar{Y} est construit de façon telle que

$$\{\theta | r_{ns}(\theta) < \chi^2_2(\alpha)\}, \quad (4.1)$$

une solution unique si, et uniquement si, $\min\{x_i, i \in s\} < 0 < \max\{x_i, i \in s\}$. La solution, si elle existe, est comprise entre $L = -1/\max\{x_i, i \in s\}$ et $U = -1/\min\{x_i, i \in s\}$. En notant que $g_2(\lambda)$ est une fonction monotone décroissante pour $\lambda \in (L, U)$, l'algorithme le plus efficace et fiable pour résoudre $g_2(\lambda) = 0$ est la méthode de bisection. La fonction `Lag1(u,d,s,mu)` fait précisément cela, où les arguments requis sont $u = (x_1, \dots, x_n)$, $ds = (d_1, \dots, d_n)$ et $mu = \bar{X}$. La sortie donne la solution de $g_2(\lambda) = 0$.

La fonction `Lag1(u,d,s,mu)` peut être utilisée conjuguée à l'approche de la pseudo-vraisemblance empirique étalonnée au moyen d'un modèle (PVEEM) de Wu et Sitter (2001) pour traiter les cas où la variable x comprend un nombre élevé de dimensions. L'approche PVEEM ne comporte qu'une seule variable de prédiction de dimension tirée d'un modèle de régression linéaire multiple et le problème connexe du multiplicateur de Lagrange est toujours unidimensionnel.

3. Échantillonnage stratifié

Soit $\{(y_{hi}, x_{hi}), i \in s_h, h = 1, \dots, H\}$ les données d'échantillon provenant d'un plan de sondage stratifié. Soit $d_{hi}^* = d_{hi} / \sum_{i \in s_h} d_{hi}$ les poids de sondage normalisés pour la strate $h, h = 1, \dots, H$. La fonction de pseudo-vraisemblance empirique sous échantillonnage stratifié définie par Wu et Rao (2004) est donnée par

$$l^{st}(d_1, \dots, d_H) = n \sum_{h=1}^H W_h \sum_{i \in s_h} d_{hi}^* \log(p_{hi}), \quad (3.1)$$

où les $W_h = N_h / N$ sont les poids de strate et n^* est la taille totale effective d'échantillon telle que définie dans Wu et Rao (2004). La valeur de n^* n'est pas nécessaire pour l'estimation ponctuelle, mais cette constante de mise à l'échelle est requise pour la construction des intervalles de confiance. Soit \bar{X} le vecteur connu des moyennes de population pour les variables auxiliaires. L'estimateur du maximum de pseudo-vraisemblance empirique de la moyenne de population $\bar{Y} = \sum_{h=1}^H W_h \bar{Y}_{hi}$ est défini comme étant $\bar{Y}_{PEL} = \sum_{h=1}^H W_h \sum_{i \in s_h} \hat{p}_{hi} y_{hi}$, où les \hat{p}_{hi} maximisent $l^{st}(d_1, \dots, d_H)$ sous l'ensemble de contraintes

$$p_{hi} > 0, \sum_{i \in s_h} p_{hi} = 1, h = 1, \dots, H$$

(3.2)

$$\sum_{i \in s_h} W_h \sum_{i \in s_h} p_{hi} x_{hi} = \bar{X}.$$

Sous échantillonnage stratifié, la principale difficulté de calcul est due au fait que la sous-normalisation des poids (c'est-à-dire $\sum_{i \in s_h} p_{hi} = 1$) a lieu au niveau de la strate, alors que les contraintes d'échantillonnage (c'est-à-dire

décrite dans Wu et Rao (2004) et ont donné de très bons résultats.

2. Échantillonnage non stratifié

Considérons une population finie constituée de N unités identifiables. Associées à la i^{e} unité sont des valeurs de la variable étudiée, y_i , et un vecteur de variables auxiliaires, x_i . Le vecteur de moyennes de population $\bar{X} = N^{-1} \sum_{i=1}^N x_i$ est connu. Soit $\{(y_i, x_i), i \in s\}$ les données d'échantillon, où s est l'ensemble d'unités sélectionnées selon un plan de sondage complexe. Soit $\pi_i = P(i \in s)$ les probabilités de sélection et $d_i = 1/\pi_i$ les poids de sondage.

L'estimateur du maximum de pseudo-vraisemblance empirique de la moyenne de population $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$ est calculé comme étant $\hat{Y}_{\text{PEL}} = \sum_{i \in s} \hat{p}_i y_i$, où les poids \hat{p}_i sont obtenus en maximisant la fonction de pseudo log-vraisemblance empirique

$$l^{ns}(\mathbf{p}) = n \sum_{i \in s} d_i^* \log(p_i) \quad (2.1)$$

sous les contraintes

$$0 < p_i < 1, \sum_{i \in s} p_i = 1 \text{ et } \sum_{i \in s} p_i x_i = \bar{X}. \quad (2.2)$$

La fonction de pseudo-vraisemblance empirique originale proposée par Chen et Sitter (1999) est $l(\mathbf{p}) = \sum_{i \in s} d_i \log(p_i)$. La fonction de pseudo-vraisemblance empirique $l^{ns}(\mathbf{p})$ donnée par (2.1) a été utilisée par Wu et Rao (2004), où les $d_i^* = d_i / \sum_{i \in s} d_i$ sont les poids de sondage normalisés et n^* est la taille effective d'échantillon. L'estimateur ponctuel $\hat{Y}_{\text{PEL}} = \sum_{i \in s} \hat{p}_i y_i$ reste le même pour l'une et l'autre version de la fonction de vraisemblance. Le développement utilisé dans $l^{ns}(\mathbf{p})$ facilite la construction des intervalles de confiance des rapports de pseudo-vraisemblance empirique.

En utilisant comme argument un multiplicateur de Lagrange standard, nous pouvons montrer que

$$\hat{p}_i = \frac{1 + \lambda_i' x_i - \bar{X}}{d_i^*} \quad \text{pour } i \in s, \quad (2.3)$$

où le multiplicateur de la Lagrange évalué vectoriellement, λ , est la solution de

$$g_1(\lambda) = \sum_{i \in s} \frac{d_i^* \lambda_i' x_i - \bar{X}}{1 + \lambda_i' x_i - \bar{X}} = 0.$$

Ici, la principale tâche de calcul consiste à trouver la solution de $g_1(\lambda) = 0$, ce qui peut se faire en utilisant la procédure de Newton-Raphson modifiée proposée par Chen et coll. (2002). La modification comprend la vérification, à chaque étape de mise à jour, que la contrainte $1 + \lambda_i' x_i - \bar{X} > 0$ (i.e., $p_i > 0$) est encore satisfaite. Sans perte de généralité, nous supposons que $\bar{X} = 0$ (sinon il

modifiée est la suivante.

Étape 0 : Soit $\lambda_0 = \mathbf{0}$. Fixer $k = 0$, $\gamma_0 = 1$ et $\varepsilon = 10^{-8}$.

Étape 1 : Calculer $\Delta_1(\lambda_k)$ et $\Delta_2(\lambda_k)$, où

$$\Delta_1(\lambda) = \sum_{i \in s} d_i^* \frac{1 + \lambda_i' x_i}{x_i} \quad \text{et} \quad \Delta_2(\lambda) = \left\{ -\sum_{i \in s} d_i^* \frac{(1 + \lambda_i' x_i)^2}{x_i^2} \Delta_1(\lambda) \right\}^{-1}$$

Si $\|\Delta_2(\lambda_k)\| < \varepsilon$, arrêter l'algorithme et donner la valeur de λ_k dans le rapport; autrement, passer à l'étape 2.

Étape 2 : Calculer $\delta_k = \gamma_k \Delta_2(\lambda_k)$. Si $1 + (\lambda_k - \delta_k)' x_i \leq 0$ pour tout i , poser que $\gamma_k = \gamma_k / 2$ et répéter l'étape 2.

Étape 3 : Poser que $\lambda_{k+1} = \lambda_k - \delta_k$, $k = k + 1$ et $\gamma_{k+1} = (\gamma_k + 1)^{-1/2}$. Passer à l'étape 1.

Dans l'algorithme original présenté par Chen et coll. (2002), l'étape 2 consiste aussi à vérifier une fonction objective duale connexe. Bien qu'elle soit nécessaire pour la preuve théorique de la convergence de l'algorithme, cette vérification n'est pas vraiment requise pour les applications pratiques.

La fonction R `Lag2(u,ds,mu)` peut être utilisée pour trouver la solution de $g_1(\lambda) = 0$ quand le vecteur de variables auxiliaires x est de dimension m et que $m \geq 2$. Quand x est univarié, une méthode de bisection extrêmement simple et stable qui sera décrite bientôt devrait être utilisée. Soit n la taille d'échantillon. Les trois arguments reçus sont la matrice de données u de dimensions $n \times m$, le vecteur de poids de sondage ds de dimension $n \times 1$ et le vecteur de moyennes de population μ de dimension $m \times 1$. La sortie de la fonction `Lag2(u,ds,mu)` donne la valeur de λ qui est la solution de $g_1(\lambda) = 0$.

La fonction `Lag2(u,ds,mu)` ne fournira pas de solution si (i) le vecteur moyen \bar{X} n'est pas un point intérieur de l'enveloppe convexe formée par $\{x_i, i \in s\}$, ou que (ii) la matrice $\sum_{i \in s} d_i^* x_i x_i'$ n'est pas de plein rang. Dans le cas (i), l'estimateur du maximum de pseudo-vraisemblance empirique n'existe pas. Ceci se produit avec une probabilité s'approchant de zéro à mesure que la taille d'échantillon n tend vers l'infini; dans le cas (ii), on peut envisager de supprimer certaines composantes des variables x de l'ensemble de contraintes (2.2) pour éliminer le problème de colinéarité.

Si la variable x est univariée, il en est de même du multiplicateur de Lagrange λ concerné. Dans ces conditions, nous devons résoudre $g_2(\lambda) = \sum_{i \in s} d_i^* x_i / (1 + \lambda x_i) = 0$ pour un scalaire λ , en supposant que $\bar{X} = 0$. Il existe

Algorithmes et codes R pour la méthode de la pseudo-vraisemblance empirique dans les sondages

Changbao Wu¹

Résumé

Nous présentons des algorithmes informatiques pour la méthode de la pseudo-vraisemblance empirique proposée récemment pour l'analyse des données d'enquête complexes. Plusieurs algorithmes essentiels pour le calcul des estimateurs du maximum de pseudo-vraisemblance empirique et la construction des intervalles de confiance des rapports de pseudo-vraisemblance empirique sont implantés au moyen des logiciels statistiques R et S-PLUS d'usage très répandu. Les codes principaux sont écrits sous la forme de fonctions R/S-PLUS et peuvent donc être utilisés directement dans les applications d'enquête et (ou) les études en simulation.

Mots clés : Intervalle de confiance; algorithme de bisection; vraisemblance empirique; procédure de Newton-Raphson; échantillonnage stratifié; échantillonnage avec probabilités inégales.

1. Introduction

L'un des grands défis que pose l'application de méthodes statistiques avancées et souvent complexes à des sondages réels est l'implantation informatique de la méthode. Souvent, des considérations pratiques obligent à rejeter des méthodes théoriquement valides et séduisantes, mais nécessitant une quantité incroyablement élevée de calculs.

La méthode de la vraisemblance empirique, proposée pour la première fois par Owen (1988), est l'un des principaux progrès réalisés en statistique au cours des 15 dernières années. Outre le fait qu'elle soit axée sur les données et qu'elle respecte les gammes de valeurs dans l'estimation et les tests, sa nature non paramétrique et discrète est particulièrement intéressante pour la résolution de problèmes en population finie. En effet, l'une de ses premières versions, appelée méthode des estimateurs « scale-load », a été utilisée en sondage par Hartley et Rao en 1968. L'étude plus récente de cette méthode dans le contexte des sondages a donné lieu à la publication d'une série de documents de recherche et suscité chez les statisticiens d'enquête un vif intérêt qui les a poussés à l'explorer plus en détail. Wu et Rao (2004) résument brièvement les faits récents concernant la méthode de la pseudo-vraisemblance empirique (PEL pour Pseudo Empirical Likelihood).

Des progrès ont également été réalisés en ce qui concerne l'élaboration d'algorithmes. Chen, Sitter et Wu (2002) ont proposé une procédure de Newton-Raphson modifiée pour calculer les estimateurs du maximum de pseudo-vraisemblance empirique sous échantillonnage non stratifié. Wu (2004a) a poursuivi la modification de la procédure afin de permettre le traitement des plans de sondage stratifiés.

Dans le présent article, nous présentons des algorithmes informatiques permettant de calculer les estimateurs du maximum de pseudo-vraisemblance empirique et de construire les intervalles de confiance des rapports de pseudo-vraisemblance empirique connexes pour des sondages complexes sous un cadre unifié, en mettant surtout l'accent sur l'implantation de ces algorithmes au moyen des logiciels R et S-PLUS. Le logiciel R, un environnement de programmation convivial compatible avec le logiciel statistiques commercial S-PLUS très répandu, intéresse de plus en plus les statisticiens. L'un des avantages de l'utilisation du logiciel R est qu'il est offert gratuitement pour la recherche et qu'il peut être téléchargé facilement à partir d'Internet. Nous espérons que le présent article comblera le fossé qui existe à l'heure actuelle entre les développements théoriques et les applications pratiques de la méthode de la pseudo-vraisemblance empirique et qu'il suscitera d'autres travaux de recherche dans ce domaine en vue de rendre l'utilisation de cette méthode entièrement pratique.

L'algorithme de calcul de l'estimateur du maximum de pseudo-vraisemblance empirique sous échantillonnage non stratifié et certaines remarques sur son implantation dans R/S-PLUS sont présentées à la section 2. L'algorithme de Wu (2004a) pour l'échantillonnage stratifié est discuté à la section 3. La construction de l'intervalle de confiance du rapport de pseudo-vraisemblance empirique, qui comprend l'établissement du profil de cette statistique, est décrite en détail à la section 4. Tous les exemples de code ou de fonction R figurent à l'annexe. Ils peuvent être téléchargés à partir de la page d'accueil personnelle de l'auteur à <http://www.stats.uwaterloo.ca/~cbwu/paper.html>. Ces codes et fonctions ont été testés lors de l'étude en simulation

clair que la matrice $r \times r$ des corrélations pour le vecteur aléatoire $\bar{\theta} = (\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \dots, \vartheta_{I_r})$, où $r \leq n$, peut s'écrire :

$$(44) \quad \text{Corr}(\bar{\theta}) = R_r(p) = \begin{pmatrix} 1 & p & \dots & p \\ p & 1 & \dots & p \\ \vdots & \vdots & \ddots & \vdots \\ p & p & \dots & 1 \end{pmatrix}.$$

Il convient de souligner que les éléments de $R_r(p)$ dans (44) dépendent uniquement des probabilités d'inclusion qui, pour toute taille d'échantillon, peuvent être calculées entièrement d'après la récurson (7) et les expressions (8) et (10). Autrement dit, elles ne dépendent d'aucun paramètre de population inconnu qu'il faut estimer ni des valeurs des variables qui doivent être mesurées sur les unités d'échantillonage.

6. Remarques finales

En théorie, l'efficacité de toute méthode d'estimation s'accroît dans une certaine mesure si l'on tenait compte explicitement de la corrélation entre les unités d'échantillonage. Il en serait certainement ainsi pour l'estimation linéaire et, dans certains cas, pour l'estimation du maximum

de vraisemblance. Par ailleurs, il convient d'insister sur le fait que $R_p(p)$ peut devenir singulière à mesure que la taille de l'échantillon n s'approche de la taille de la population N ; il en est ainsi pour l'EAS ($R_N(-1/(N-1))$, ainsi que pour l'échantillonage sans remise en général. Par conséquent, numériquement, nombre de méthodes d'estimation qui s'appuient sur l'inverse ou le déterminant de R plutôt que sur la matrice des corrélations proprement dite pourraient également bénéficier du remplacement de l'hypothèse simplifiante d'indépendance entre les observations quand la taille de l'échantillon est grande relativement à la taille de la population. Les cas où cela est possible se produisent à certaines étapes dans l'échantillonnage à plusieurs degrés (par exemple nombre de ménages dans un îlot) et dans de grandes enquêtes à l'échelle du pays.

$$(41) \quad \begin{aligned} \text{Cov}(\vartheta_{I_1}, \vartheta_{I_r}) &= \frac{1}{N} \sum_{j=1}^{r-1} \sum_{i=j+1}^r \frac{u(u-1)}{n(n-1)} \left(\frac{1}{N} \sum_{i=1}^n \left(\frac{N}{n} \right)^2 \right. \\ &\quad \left. - \frac{n(n-1)}{N^2} - \frac{N^2}{n^2} \right), \end{aligned}$$

tandis que dans le cas sans remise, on peut voir que la covariance est égale à

$$(42) \quad \begin{aligned} \text{Cov}(\vartheta_{I_1}, \vartheta_{I_r}) &= \frac{1}{N} \sum_{j=1}^{r-1} \sum_{i=j+1}^r \frac{u(u-1)}{N(N-1)} \left(\frac{1}{N} \sum_{i=1}^n \left(\frac{N}{n} \right)^2 \right. \\ &\quad \left. - \frac{n(n-1)}{N^2} - \frac{N^2}{n^2} \right). \end{aligned}$$

Il convient de souligner que, dans le cas de l'EAS, que le tirage se fasse avec ou sans remise, les coefficients de corrélation sont donnés par

$$(43) \quad \text{Corr}(\vartheta_{I_1}, \vartheta_{I_r}) = \frac{-1}{N-1}.$$

Indépendamment de la taille de l'échantillon. De surcroît, nous savons que, à mesure que la valeur de n s'approche de N dans l'échantillonnage sans remise, $\pi_{(n)}^{I_1}$ et $\pi_{(n)}^{I_r}$ s'approchent l'une et l'autre de l'unité. En particulier, quand $n = N$, les valeurs des expressions (31) et (40) deviennent nulles.

5. La matrice des corrélations des unités d'échantillonnage

Dès que l'on se rend compte qu'aucune des expressions (28), (31) et (40) ne dépend d'aucun des indices arbitraires utilisés pour différencier les unités de population, il devient

Tandis que (31) donne

$$V(\vartheta_{I_j}) = \left(\frac{1}{N} \sum_{l=1}^n \left(\frac{N}{n} \right)_2 \right) \left[1 - \left(\frac{n}{N} \sum_{l=1}^n \left(\frac{N}{n} \right)_2 \right) \right] \quad (33)$$

4.3 La covariance entre les unités d'échantillonnage

Afin d'établir la covariance entre les diverses unités d'échantillonnage, nous recourons à une simple extension

de (25),

$$\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = \text{Cov}_{I_j, I_k}[E(\vartheta_{I_j} | I_j), E(\vartheta_{I_k} | I_k)] + E_{I_j, I_k}[\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j, I_k)]. \quad (34)$$

Dans ce cas, nous savons que

$$E(\vartheta_{I_j} | I_j = I) = \pi_{(n)}^I \quad (35)$$

et

$$E(\vartheta_{I_j}, \vartheta_{I_k} | I_j = I, I_k = J) = \pi_{(n)}^{IJ} \quad (36)$$

tandis qu'il est facile de voir que la covariance entre crochets dans le deuxième membre de (34) est égale à

$$\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j = I, I_k = J) = \pi_{(n)}^{IJ} - \pi_{(n)}^I \pi_{(n)}^J. \quad (37)$$

À partir de (35) et (36), nous obtenons

$$\text{Cov}_{I_j, I_k}[E(\vartheta_{I_j} | I_j), E(\vartheta_{I_k} | I_k)] = E_{I_j, I_k}(\pi_{(n)}^I \pi_{(n)}^J) - E_{I_j, I_k}(\pi_{(n)}^{IJ}) \quad (38)$$

tandis que, de (37), nous obtenons

$$E_{I_j, I_k}[\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j, I_k)] = E_{I_j, I_k}(\pi_{(n)}^{IJ}) - E_{I_j, I_k}(\pi_{(n)}^I \pi_{(n)}^J). \quad (39)$$

Enfin, en additionnant ces deux dernières expressions,

nous obtenons la covariance souhaitée

$$\text{Cov}(\vartheta_{I_j}, \vartheta_{I_k})$$

$$= E_{I_j, I_k}(\pi_{(n)}^{IJ}) - [E_{I_j, I_k}(\pi_{(n)}^I)] [E_{I_k}(\pi_{(n)}^J)]$$

$$= \frac{1}{n(n-1)} \sum_{l=1}^n \sum_{l' \neq l}^n (\pi_{(n)}^{ll'})_2 - \left(\frac{n}{N} \sum_{l=1}^n (\pi_{(n)}^l)_2 \right) \left(\frac{n}{N} \sum_{l=1}^n (\pi_{(n)}^l)_2 \right) \quad (40)$$

Dans le cas de l'EAS/WR, (40) donne

$$E(\vartheta_{I_j}) = \frac{1}{N} \sum_{l=1}^n (\pi_{(n)}^l)_2 = \frac{1}{N} \left(\frac{n}{N} \right)_2 = \frac{n}{N} \cdot \frac{n}{N} \quad (32)$$

devient

d'un exemple, considérons l'EAS. L'expression (30)

$$V(\vartheta_{I_j}) = E(\pi_{(n)}^{I_j}) - E_2(\pi_{(n)}^{I_j}) [1 - E(\pi_{(n)}^{I_j})] \quad (31)$$

Donc, la variance est donnée par

$$E[V(\vartheta_{I_j} | I_j)] = E[\pi_{(n)}^{I_j} (1 - \pi_{(n)}^{I_j})] = E[\pi_{(n)}^{I_j}] - [E(\pi_{(n)}^{I_j})]^2$$

et

$$V[E(\vartheta_{I_j} | I_j)] = V(\pi_{(n)}^{I_j}) = E[(\pi_{(n)}^{I_j})^2] - [E(\pi_{(n)}^{I_j})]^2$$

donc nous tirons

$$E(\vartheta_{I_j} | I_j) = \pi_{(n)}^{I_j} \text{ et } V(\vartheta_{I_j} | I_j) = \pi_{(n)}^{I_j} (1 - \pi_{(n)}^{I_j})$$

noter que

En utilisant de nouveau (25), nous commençons par

$$E(\vartheta_{I_j}) = \frac{1}{N} \sum_{l=1}^n \pi_{(n)}^l (d_{(I_j)}^l)_2 = \frac{1}{N} \sum_{l=1}^n \pi_{(n)}^l (d_{(I_j)}^l)_2 \quad (30)$$

et, par conséquent,

$$P(\vartheta_{I_j} = x) = \frac{1}{N} \sum_{l=1}^n \pi_{(n)}^l \sum_{k=1}^K (d_{(I_j)}^k)^x (1 - d_{(I_j)}^k)^{1-x} \quad (29)$$

par

Dans ce cas, la fonction de probabilité de ϑ_{I_j} est donnée

4.2 Moyenne et variance pour l'échantillonnage sans remise

$$V(\vartheta_{I_j}) = \sum_{l=1}^n \frac{1}{N} n \frac{N}{1} \left(1 + (n-1) \frac{1}{N} \frac{N}{1} \sum_{l=1}^n \frac{N}{1} \right) = n \frac{N}{1} \left(1 - \frac{1}{N} \right)$$

Tandis que (28) donne

$$E(\vartheta_{I_j}) = \frac{1}{N} \sum_{l=1}^n \pi_{(n)}^l (w_{(n)}^l)_2 = \frac{1}{N} \sum_{l=1}^n \left(\frac{N}{n} \right)_2 = \left(\frac{N}{n} \right)_2 = \frac{n}{N}.$$

Pour le cas de l'EAS, l'équation (24) qui précède donne

$$= \sum_{l=1}^n n p_{I_j}^2 \left(1 + (n-1) p_{I_j} - \sum_{N}^f n p_{I_j}^2 \right) \quad (28)$$

$$= n [E_{I_j}(p_{I_j}^2) - E_{I_j}(p_{I_j}^2)] + n_2 [E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})]$$

$$V(\vartheta_{I_j})$$

4. Les deux premiers moments des unités d'échantillonnage

Après avoir établi les moments de premier et de deuxième ordre du vecteur $\vec{\vartheta}$, il nous est possible de déterminer les moments correspondants de sous-vecteurs de diverses tailles et dont les composantes sont sélectionnées aléatoirement, c'est-à-dire l'échantillon. À cette fin, définissons les variables aléatoires $\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \dots, \vartheta_{I_r}$, où r représente le nombre d'unités de population différentes dans l'échantillon et dont les indices $I_k, 1 \leq k \leq r \leq n$, peuvent prendre la valeur I avec la probabilité $\pi_{(n)}^I$. Autrement dit, sous les conditions susmentionnées, nous sommes en présence d'un jeu de variables aléatoires dont les indices sont eux-mêmes aléatoires.

4.1 Moyenne et variance pour l'échantillonnage avec remise

Dans ce cas, la fonction de probabilité de ϑ_{I_1} est donnée par

$$P(\vartheta_{I_1} = x) = \sum_{I=1}^I p_I P(\vartheta_I = x) \quad (23)$$

$$= \sum_{I=1}^I p_I \left(\frac{x}{n} \right)^n p_I^{n-x}.$$

Les deux premiers moments peuvent aussi être obtenus par la voie d'un argument conditionnel. La moyenne de sa distribution est donnée par

$$E(\vartheta_{I_1}) = \sum_{I=1}^I p_I E(\vartheta_I) = \sum_{I=1}^I n p_I p_I = n \sum_{I=1}^I p_I^2. \quad (24)$$

À son tour, sa variance est calculée à l'aide de la formule bien connue

$$V(\vartheta_{I_1}) = V_{I_1}[E(\vartheta_{I_1} | I_1)] + E_{I_1}[V(\vartheta_{I_1} | I_1)]. \quad (25)$$

Dans ce cas, nous avons

$$E(\vartheta_{I_1} | I_1 = I) = n p_I \quad \text{et} \quad V(\vartheta_{I_1} | I_1 = I) = n p_I (1 - p_I). \quad (26)$$

Donc,

$$V_{I_1}[E(\vartheta_{I_1} | I_1)] = V_{I_1}(n p_I) \\ = n^2 [E_{I_1}(p_I^2) - E_{I_1}^2(p_I)], \\ E_{I_1}[V(\vartheta_{I_1} | I_1)] = n^2 p_I [1 - p_I] \\ = n [E_{I_1}(p_I) - E_{I_1}^2(p_I)] \quad (27)$$

et, par conséquent

que la matrice C est singulière. Sinon, il est impossible que

l'équation (20) soit satisfait.

En fait, il est possible de prouver que la somme des éléments de toute ligne (ou colonne) de C doit être nulle, ce qui est un énoncé plus ferme. Sachant que la covariance entre une variable aléatoire et une constante est nulle, nous

obtenons

$$0 = \text{Cov}(\vartheta_1, n) = \text{Cov}(\vartheta_1, \vartheta_1 + \vartheta_2 + \vartheta_3 + \dots + \vartheta_N) \\ = C_{I1} + C_{I2} + \dots + C_{IN} \\ = \text{Var}(\vartheta_1) + \sum_{j \neq 1}^j \text{Cov}(\vartheta_1, \vartheta_j). \quad (21)$$

Nous avons donc prouvé que, dans le cas de l'échantillonnage sans remise, (22.1) est vérifiée.

Le même énoncé peut être prouvé algébriquement en notant que

$$\sum_{I=1}^I \pi_{(n)}^I = \pi_{(n)}^I \sum_{I=1}^I \pi_{(n)}^I \\ = (n - 1) \pi_{(n)}^I,$$

ce qui est évident si l'on se rend compte que la probabilité conditionnelle concernée représente la probabilité que l'unité de population I entre dans un échantillon de taille $n - 1$ pour lequel s'applique aussi l'expression (19). En outre, en utilisant de nouveau (19), notons que

$$\sum_{j \neq I}^j \pi_{(n)}^j = (n - \pi_{(n)}^I),$$

et, par conséquent,

$$0 = \pi_{(n)}^I (1 - \pi_{(n)}^I) + \sum_{j \neq I}^j (\pi_{(n)}^I - \pi_{(n)}^j) \pi_{(n)}^j$$

$$= \pi_{(n)}^I - (\pi_{(n)}^I)^2 + (n - 1) \pi_{(n)}^I - \pi_{(n)}^I (n - \pi_{(n)}^I).$$

Pour l'échantillonnage avec remise, (21) implique que :

$$0 = n p_I q_I + \sum_{j \neq I}^j (n(n - 1) p_I p_j - n^2 p_I p_j) \\ = n p_I q_I - n p_I \sum_{j \neq I}^j p_j \quad (22.2)$$

condition qui, on le voit directement, s'applique.

En tout cas, l'incidence la plus importante des résultats susmentionnés est que, indépendamment du plan d'échantillonnage, la matrice de corrélation des variables aléatoires de population $\vartheta_1, \vartheta_2, \vartheta_3, \dots, \vartheta_N$ est singulière. En ce qui concerne les situations pratiques décrites dans l'introduction, la conséquence la plus importante tient principalement au fait que l'inverse de la matrice des covariances est utilisée dans de nombreuses méthodes d'ajustement et

d'estimation de modèles.

composantes de ce vecteur peuvent être exprimées sous forme du produit de $\bar{\vartheta}_{I^0}$ par un vecteur dont les composantes sont toutes égales à un. Autrement dit,

$$(13) \quad \bar{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ \vdots \\ \vartheta_N \end{pmatrix} = \begin{pmatrix} \bar{\vartheta}_1^T \bar{1} \\ \bar{\vartheta}_2^T \bar{1} \\ \bar{\vartheta}_3^T \bar{1} \\ \vdots \\ \bar{\vartheta}_N^T \bar{1} \end{pmatrix}.$$

Certaines propriétés distributionnelles de ces sommes peuvent alors être obtenues directement d'après celles des lignes ou des colonnes de la matrice V .

Par exemple, leurs valeurs attendues sont données par

$$E(\vartheta_j) = E(\bar{\vartheta}_{I^0}^T \bar{1}) = E\left(\sum_{k=1}^K \vartheta_{jk}\right)$$

$$= \sum_{n=1}^N \pi_{(1)}^T d_{(k)}^I = \pi_{(1)}^T + \sum_{n=2}^K (\pi_{(k)}^T - \pi_{(k-1)}^T). \quad (14)$$

De (1,2), nous tirons la restriction non stochastique :

$$\bar{1}^T \bar{\vartheta} = \vartheta_1 + \vartheta_2 + \vartheta_3 + \dots + \vartheta_N = n. \quad (15)$$

De (14) et (15) découlent directement les propositions bien connues (16) et (17),

$$E[\bar{\vartheta}^T] = (\pi_{(1)}^T, \pi_{(2)}^T, \pi_{(3)}^T, \dots, \pi_{(N)}^T) \quad (16)$$

$$\pi_{(n)}^T + \pi_{(n)}^2 + \pi_{(n)}^3 + \dots + \pi_{(n)}^N = n. \quad (17)$$

Pour les moments de deuxième ordre, nous obtenons

$$\text{Cov}(\vartheta_I, \vartheta_J) = \text{Cov}(\bar{1}^T \bar{\vartheta}_{I^0}, \bar{1}^T \bar{\vartheta}_{J^0})$$

$$= -\bar{1}^T \text{Cov}(\bar{\vartheta}_{I^0}, \bar{\vartheta}_{J^0}) \bar{1} = -\sum_{k=1}^K d_{(k)}^I d_{(k)}^J$$

$$= \begin{cases} -nd^I d^J & \text{WR} \\ \pi_{(n)}^T - \pi_{(n)}^I \pi_{(n)}^J & \text{WOR}, \end{cases}$$

(18)

qui indique clairement que la covariance n'est jamais positive. À leur tour, les variances sont données par

$$\text{Var}(\vartheta_I) = \text{Var}(\bar{1}^T \bar{\vartheta}_{I^0}) = \bar{1}^T \text{Cov}(\bar{\vartheta}_{I^0}, \bar{\vartheta}_{I^0}) \bar{1}$$

$$= \begin{cases} nd^I d^I & \text{WR} \\ \pi_{(n)}^T (1 - \pi_{(n)}^I) & \text{WOR}. \end{cases} \quad (19)$$

Une autre conséquence importante de (15) concerne les moments de deuxième ordre du vecteur stochastique $\bar{\vartheta}$.

$$0 = \text{Var}(n) = \text{Var}(\bar{1}^T \bar{\vartheta}) = \bar{1}^T \text{Cov}(\bar{\vartheta}, \bar{\vartheta}) \bar{1} = \bar{1}^T C \bar{1}. \quad (20)$$

De toute évidence, les éléments diagonaux de la matrice C , la matrice des covariances de $\bar{\vartheta}$, ne sont pas tous nuls. Par conséquent, le tirage aléatoire d'un échantillon de taille fixe introduit dans les unités de population une dépendance qui donne lieu à des covariances non nulles sous-entendant

Considérons maintenant les vecteurs de ligne $\bar{\vartheta}_{I^0}$. Alors, pour la matrice des covariances entre diverses lignes, nous obtenons

$$(10.2) \quad \sum_{j=1}^J \left(\frac{N(N-1)}{n-f} + \frac{N(N-1)}{n-f} \right) = \frac{N(N-1)}{n(N-1)}.$$

$$\pi_{(n)}^T = \sum_{j=1}^J \left(d_{(j)}^I d_{(j)}^T + d_{(k)}^I d_{(k)}^T + \sum_{n \neq j < k}^J d_{(j)}^I d_{(k)}^T \right) \text{ où } j \neq I$$

$$(8.2) \quad \pi_{(n)}^T = \frac{N}{n}$$

$$(7.2) \quad d_{(k)}^I = \frac{N}{1} \text{ quand } k \geq 1$$

les expressions (7,2), (8,2) et (10,2).

Dans le cas de l'EAS/WOR, nous obtenons, à la place,

$$(10.1) \quad \pi_{(n)}^T = \sum_{j=1}^J \left(d_{(j)}^I d_{(j)}^T + d_{(k)}^I d_{(k)}^T + \sum_{n \neq j < k}^J d_{(j)}^I d_{(k)}^T \right) = \frac{N^2}{n(N-1)} + \frac{N^2}{n-f} + \frac{N^2}{n(N-1)}.$$

Dans le cas de l'échantillonnage sans remise, la matrice des covariances susmentionnée devient

$$\text{Cov}(\bar{\vartheta}_{I^0}, \bar{\vartheta}_{J^0}) = \begin{bmatrix} d^I d^I & 0 & 0 & 0 \\ 0 & d^I d^I & 0 & 0 \\ 0 & 0 & d^I d^I & 0 \\ 0 & 0 & 0 & d^I d^I \end{bmatrix} \quad (12.1)$$

En cas d'échantillonnage avec remise où, par conséquent, $d^I = d^I \forall j = 1, \dots, n$, la matrice des covariances pour le I^e vecteur de ligne est donnée par

$$\text{Cov}(\bar{\vartheta}_{I^0}, \bar{\vartheta}_{J^0}) = \begin{bmatrix} d_{(1)}^I d_{(1)}^T & 0 & 0 & 0 \\ 0 & d_{(2)}^I d_{(2)}^T & 0 & 0 \\ 0 & 0 & d_{(3)}^I d_{(3)}^T & 0 \\ 0 & 0 & 0 & d_{(n)}^I d_{(n)}^T \end{bmatrix} \quad (11)$$

Soit $\bar{\vartheta}$ le vecteur de dimension N qui résulte de l'addition des colonnes de V . De toute évidence, les

$$(12.2) \quad \text{Cov}(\bar{\vartheta}_{I^0}, \bar{\vartheta}_{J^0}) = \begin{bmatrix} d_{(1)}^I (1 - d_{(1)}^I) & -d_{(1)}^I d_{(2)}^I & \dots & -d_{(1)}^I d_{(n)}^I \\ -d_{(2)}^I d_{(1)}^I & d_{(2)}^I (1 - d_{(2)}^I) & \dots & -d_{(2)}^I d_{(n)}^I \\ \vdots & \vdots & \ddots & \vdots \\ -d_{(n)}^I d_{(1)}^I & -d_{(n)}^I d_{(2)}^I & \dots & d_{(n)}^I (1 - d_{(n)}^I) \end{bmatrix}.$$

ligne l ne peut prendre que deux valeurs : un, si la l^{e} unité est tirée à un certain degré, ou zéro, autrement, ce qui nous ramène au cas de Bernoulli.

Nous pouvons former des sous-ensembles disjoints de lignes conformément à divers critères. Par exemple, si nous regroupons les lignes en fonction de leur voisinage spatial, nous pourrions parler de grappes ou d'unités primaires d'échantillonnage. Si nous fondons le groupement sur un ou plusieurs indicateurs statistiques, nous utilisons habituellement le terme de strate.

Définissons maintenant les probabilités d'inclusion comme étant

$$\pi_{(k)}^l = P(\text{unité de population } l \text{ dans l'échantillon de taille } k) \\ = 0 \text{ si } k = 0, \\ \text{Notons que } \pi_{(n)}^l = \pi_l, \text{ habituellement notée probabilité d'inclusion de l'unité } l.$$

Représentons maintenant par $\bar{v}_{o_j}^l$ la j^{e} colonne et par $\bar{v}_{o_i}^l$ la i^{e} ligne de la matrice V . Par conséquent, en nous basant sur l'expression suivante,

$$f(\bar{v}_{o_1}^l, \bar{v}_{o_2}^l, \bar{v}_{o_3}^l, \dots, \bar{v}_{o_n}^l) = f(\bar{v}_{o_1}^l | \bar{v}_{o_2}^l, \bar{v}_{o_3}^l, \dots, \bar{v}_{o_{n-1}}^l) f(\bar{v}_{o_2}^l | \bar{v}_{o_3}^l, \bar{v}_{o_4}^l, \dots, \bar{v}_{o_{n-1}}^l) \dots f(\bar{v}_{o_{n-1}}^l | \bar{v}_{o_n}^l), \quad (3)$$

nous pouvons écrire la fonction de probabilité conjointe des éléments de V sous la forme :

$$f(\bar{v}_{o_1}^l, \bar{v}_{o_2}^l, \bar{v}_{o_3}^l, \dots, \bar{v}_{o_n}^l) = \prod_{i=1}^n \left[\prod_{j=1}^n (\pi_{(k-1)}^l)^{\bar{v}_{o_j}^l} (1 - \pi_{(k-1)}^l)^{1 - \bar{v}_{o_j}^l} \right] \prod_{k=1}^n \left[\prod_{j=1}^n (d_{(k)}^l)^{\bar{v}_{o_j}^l} \right] \quad (4)$$

sachant que

$$\sum_{l=1}^I \bar{v}_{o_k}^l = 1, k = 1, \dots, n \text{ et } \sum_{k=1}^n \bar{v}_{o_k}^l \leq \begin{cases} 1, \text{WOR} \\ n, \text{WR} \end{cases}$$

et ici $d_{(k)}^l$, définie comme étant $d_{(k)}^l = (\pi_{(k)}^l - \pi_{(k-1)}^l)$, représente la probabilité que l'unité de population l soit incluse dans l'échantillon lors du k^{e} tirage. La fonction susmentionnée est utile pour le calcul de la probabilité de tout échantillon ordonné de taille n . Manifestement, si l'on peut ignorer l'ordre d'inclusion, on obtiendra la probabilité d'un échantillon donné en ajoutant les $n!$ valeurs obtenues au moyen de (4).

3. Les incidences de l'échantillonnage sur les propriétés stochastiques des unités de population

Conséquemment,

$$E(\bar{v}_{o_k}^l) = d_{(k)}^l = (\pi_{(k)}^l - \pi_{(k-1)}^l) \quad (5)$$

tion (7) qui suit.

$$d_{(k)}^l = \begin{cases} d_{(k-1)}^l \sum_{j \neq l}^f \frac{1 - d_{(k-1)}^j}{d_{(k-1)}^l} & \text{si } k > 1, \\ d_l & \text{si } k = 1 \end{cases} \quad (7)$$

Il convient de souligner que (7) nous permet de calculer les probabilités souhaitées à deux moments distincts : en premier lieu, quand aucun tirage n'a effectivement eu lieu, l'ensemble de la population et, en deuxième lieu, quand on connaît le résultat du tirage précédent, moment auquel la probabilité que la j^{e} unité de population, disons, entre dans l'échantillon est égale à 1 et toutes les autres probabilités pour ce tirage sont nulles. Par conséquent, du moins en théorie, nous pouvons calculer l'inverse des facteurs dits d'expansion ou poids pour l'échantillonnage à un degré, ou étape par étape pour l'échantillonnage à plusieurs degrés.

De toute évidence,

$$\pi_{(n)}^l = \sum_{k=1}^n d_{(k)}^l. \quad (8)$$

Si nous définissons les probabilités d'inclusion conjointes comme étant

$$\pi_{(k)}^l = P \left(\begin{array}{c} \text{unités de population } l \text{ et} \\ \text{ } f \text{ dans l'échantillon de taille } k \end{array} \right), \quad (9)$$

alors nous savons qu'elles peuvent également être calculées comme suit :

$$\pi_{(n)}^l = \sum_{j=1}^{f-1} \left(d_{(f)}^l d_{(j)}^l + \sum_{k < j}^n d_{(k)}^l d_{(j)}^l + \sum_{k < j}^n d_{(k)}^l d_{(j)}^l \right). \quad (10)$$

Par exemple, dans le cas de l'échantillonnage aléatoire simple avec remise (EAS/WR), les expressions (7), (8) et (10) donnent lieu à (7.1), (8.1) et (10.1),

$$d_{(k)}^l = \frac{N}{I} \text{ quand } k \geq 1 \quad (7.1)$$

$$\pi_{(n)}^l = \frac{N}{n} \quad (8.1)$$

Structure de corrélation des unités d'échantillonnage

Alfredo Bustos¹

Résumé

Nous expliquons dans cet article certaines propriétés distributionnelles des unités d'échantillonnage qui ne sont habituellement pas décrites dans la documentation, notamment leur structure de corrélation et le fait que celle-ci ne dépend pas d'indices de population attribués arbitrairement. Ces propriétés importent pour plusieurs méthodes d'estimation, dont l'efficacité serait améliorée si on les mentionnait explicitement.

Mots clés : Recensement; enquête; échantillonnage; unités d'échantillonnage; fonction de probabilité; moyenne; covariance.

1. Introduction

Ces dernières années, la réalisation des recensements de la population et des ménages tels que nous les connaissons est devenue plus ardue pour plusieurs raisons. Par conséquent, d'autres moyens de recueillir plus fréquemment l'information requise pour la production de statistiques aux niveaux local, provincial et national ont été proposés. De grandes enquêtes nationales continues, notamment celles

appelées recensement continu, réalisées auprès d'échantillons de grande taille selon des plans d'enquête complexes, sont envisagées. Cependant, afin de produire des résultats au niveau local comparables à ceux d'un recensement, il faut mettre au point diverses méthodes d'estimation et de validation, ainsi que, dans certains cas, d'imputation et améliorer leur efficacité. Un moyen d'accroître l'efficacité consiste à tenir compte de toute l'information pertinente disponible. Naturellement, cela englobe les propriétés stochastiques des

unités d'échantillonnage. Dans la suite de l'exposé, en partant de principes fondamentaux, nous dérivons une forme générale explicite de la fonction de probabilité d'un échantillon ordonné. Nous montrons aussi comment on peut calculer cette fonction, ainsi que les probabilités d'inclusion. Enfin, nous donnons une forme générale de la matrice des corrélations des unités d'échantillonnage qui ne dépend que des probabilités d'inclusion, de sorte qu'il soit possible d'améliorer les méthodes d'estimation linéaires et du maximum de vraisemblance.

2. Le modèle de base

Le modèle de base dont nous partons représente le tirage séquentiel aléatoire de n unités à partir d'une population U

formée de N de ces unités et peut être énoncé comme suit. Soit N et n deux constantes positives telles que $n \leq N$, et soit V une matrice de dimensions $N \times n$, dont les composantes sont distribuées chacune comme des variables aléatoires de Bernoulli avec, éventuellement, des paramètres différents. Alors,

$$V^{N \times n} = \begin{bmatrix} \vartheta_{11} & \vartheta_{12} & \vartheta_{13} & \dots & \vartheta_{1n} \\ \vartheta_{21} & \vartheta_{22} & \vartheta_{23} & \dots & \vartheta_{2n} \\ \vartheta_{31} & \vartheta_{32} & \vartheta_{33} & \dots & \vartheta_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vartheta_{N1} & \vartheta_{N2} & \vartheta_{N3} & \dots & \vartheta_{Nn} \end{bmatrix} \quad (1.1)$$

Fait aussi partie du modèle la contrainte voulant que la somme des éléments de chaque colonne de V soit égale à l'unité. Autrement dit, nous exigeons que la condition

$$\sum_{i=1}^N \vartheta_{ik} = 1, \text{ for } k = 1, \dots, n \quad (1.2)$$

soit satisfaite.

Cette condition est nécessaire, parce que si le j^{e} tirage donne lieu à la sélection de l'unité de population I , alors l'élément (I, j) prend la valeur de un, tandis que tous les autres éléments de la colonne j sont nuls. Notons que cela équivaut à imposer une contrainte non stochastique au comportement de toutes les composantes de la i^{e} colonne de V , indépendamment du plan d'échantillonnage. Par conséquent, les éléments appartenant à une même colonne ne se comportent pas de façon indépendante.

Lorsque l'échantillonnage a lieu avec remise (WR pour *with replacement*), la somme des éléments de la i^{e} ligne de la matrice susmentionnée suit une loi binomiale (n, p_i) , la matrice susmentionnée est indépendante puisque la distribution de chaque colonne est indépendante de celle des autres. Par ailleurs, si l'échantillonnage se fait sans remise (WOR pour *without replacement*), le total de la

Holt, D., et Elliott, D. (1991). Methods of weighting for unit non-response. *The Statistician*, 40, 333-342.

Little, R.J. (1986). Survey nonresponse adjustment for estimate of means. *Revue Internationale de Statistique*, 54, 139-157.

Little, R.J., et Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Little, R.J., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage? *Techniques d'enquête*, 31, 175-183.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Le modèle de non-réponse final choisi s'ajustait dans une mesure raisonnable aux données de l'EPA pour la plupart des mois considérés, selon le test d'adéquation de Hosmer-Lemeshow. Néanmoins, on a utilisé la méthode des "scores" de Little (1986) pour obtenir une certaine robustesse contre des défauts non détectés du modèle. On a d'abord utilisé le modèle de non-réponse logistique décrit ci-dessus pour obtenir une probabilité de réponse estimée pour chaque ménage échantillonné, puis on a divisé l'échantillon en une cinquantaine de classes homogènes par rapport à cette probabilité de réponse estimée en utilisant l'algorithme de classification mis en œuvre dans la procédure FASTCLUS de SAS. On a pu obtenir ce nombre élevé de classes grâce à la grande taille de l'échantillon de l'EPA. On l'a choisi de manière à réduire le biais de non-réponse non seulement au niveau de la population, mais également pour les plus petits domaines. On a simplement calculé l'ajustement de poids pour la non-réponse d'un ménage répondant à un sein d'une classe donnée c en utilisant l'inverse du taux de réponse non pondéré au sein de la classe c . Un seuil pour l'ajustement de poids pour la non-réponse a été fixé à 2,5 pour contrôler la variance due à la non-réponse de l'estimateur aux poids ajustés pour la non-réponse. Il n'a fallu appliquer ce seuil que pour un nombre très petit de classes, soit celles qui présentaient les plus faibles probabilités de réponse estimées. Sans ce seuil, on aurait observé à l'occasion des ajustements de poids pour la non-réponse se situant autour de 4.

On a envisagé un autre modèle de non-réponse dans lequel on a modélisé la probabilité de réponse d'un ménage k comme le produit de la probabilité de joindre le ménage k par la probabilité de réponse de ce ménage, étant donné qu'on l'a joint. Ces deux dernières probabilités ont été modélisées séparément. Bien que ce modèle semble présenter une meilleure approximation de la réalité et qu'il ait donné des résultats légèrement supérieurs (en ce sens qu'il expliquait mieux la non-réponse), on n'a pas jugé l'amélioration suffisante pour ajouter cette complexité à la méthode d'ajustement de la non-réponse. Ce modèle pourrait cependant faire l'objet d'une étude plus approfondie.

4. Conclusion

Une contribution importante de cet article est qu'il faille considérer les renseignements sur le PCD comme aléatoires lorsqu'on les utilise dans un modèle de non-réponse. Nous avons ensuite montré que l'utilisation de ces renseignements pour traiter la non-réponse totale au moyen d'un ajustement de poids n'introduisait ni biais ni composante additionnelle de variance dans les estimations de totaux de population lorsque le modèle de non-réponse est bien spécifié. En outre, nous avons soutenu que si les renseignements sur le PCD étaient associés aux variables d'intérêt et à la non-réponse, leur utilisation avait alors tendance à réduire le biais de non-réponse lorsque le mécanisme de non-réponse dépend directement des variables d'intérêt. Enfin, au moyen de l'exemple de l'EPA, nous avons montré que ces renseignements pouvaient être utiles pour composer avec la non-réponse totale à une grande enquête.

Remerciements

L'estimateur de réponse complétée que nous avons considéré est l'estimateur de Horvitz-Thompson. Nos conclusions seraient restées les mêmes si nous avions utilisé plutôt un estimateur par la régression généralisée. Nous avons choisi l'estimateur de Horvitz-Thompson pour sa simplicité et parce qu'il était suffisant pour démontrer l'essentiel de notre exposé.

Je tiens à remercier les membres du Comité consultatif pour les méthodes statistiques de Statistique Canada pour avoir soulevé des questions concernant l'application de la méthode proposée à l'Enquête sur la population active du Canada et, en particulier, J.N.K. Rao et Chris Skinner pour leurs précieuses observations à la suite de la présentation au Comité. Je tiens aussi à remercier sincèrement le rédacteur associé pour ses observations et ses suggestions. Elles se sont avérées très utiles et ont permis d'améliorer la clarté de l'article. Enfin, je suis très reconnaissant à Asma Alavi et Cynthia Bocci de Statistique Canada pour avoir mis au point les programmes informatiques ayant servi à analyser les données de l'Enquête sur la population active du Canada.

Bibliographie

- Alavi, A., et Beaumont, J.-F. (2004). Nonresponse adjustment plans for the Labour Force Survey. Rapport technique présenté au Comité consultatif sur les méthodes statistiques, Statistique Canada, 2-3 mai 2004.
- Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B*, 67, 445-458.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue internationale de Statistique*, 51, 279-292.
- Couper, M., et Lyberg, L. (2005). The use of paradata in survey research. *Bulletin of the International Statistical Institute* (à paraître).
- Ekholm, A., et Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 7, 325-337.
- Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B., et Lindeyer, J. (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Statistique Canada, Catalogue numéro 71-526.
- Statistique Canada, N° 12-001-XPB au catalogue

L'estimation des paramètres du modèle de non-réponse sur la variance d'un estimateur d'un total de population.

3. L'exemple de l'Enquête sur la population active du Canada

L'objet de cet exemple n'est pas de présenter en détail notre analyse des données de l'Enquête sur la population active (EPA) du Canada, mais simplement de décrire certains aspects liés au choix du modèle de non-réponse et à l'estimation des probabilités de réponse. Dans cette optique, nous formulons ensuite nos principales conclusions. On trouvera dans Alavi et Beaumont (2004) des renseignements plus détaillés sur les résultats de l'analyse des données de l'EPA et sur la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

L'EPA est une enquête mensuelle menée selon un plan d'échantillonnage stratifié à plusieurs degrés (Gambino, Singh, Dufour, Kennedy et Lindeyer 1998). Les renseignements utilisés pour construire le plan de sondage et pour prélever un échantillon de logements sont essentiellement géographiques. L'échantillon est divisé en six groupes de renouvellement représentatifs et chaque logement échantillonné reste dans l'échantillon pendant six mois consécutifs. Un groupe de renouvellement contient les logements dont les membres sont interviewés pour la première fois; un autre, ceux dont les membres sont interviewés pour la deuxième fois, et ainsi de suite. C'est pourquoi, dans cinq groupes de renouvellement sur six, on trouve d'un mois à l'autre les mêmes logements échantillonnés. On utilise l'interview assistée par ordinateur pour recueillir les données d'enquête sur chaque membre des ménages sélectionnés. Ce mode de collecte permet d'obtenir une grande quantité de renseignements sur le PCD auprès des ménages répondants et non répondants.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

à la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

L'EPA est une enquête mensuelle menée selon un plan d'échantillonnage stratifié à plusieurs degrés (Gambino, Singh, Dufour, Kennedy et Lindeyer 1998). Les renseignements utilisés pour construire le plan de sondage et pour prélever un échantillon de logements sont essentiellement géographiques. L'échantillon est divisé en six groupes de renouvellement représentatifs et chaque logement échantillonné reste dans l'échantillon pendant six mois consécutifs. Un groupe de renouvellement contient les logements dont les membres sont interviewés pour la première fois; un autre, ceux dont les membres sont interviewés pour la deuxième fois, et ainsi de suite. C'est pourquoi, dans cinq groupes de renouvellement sur six, on trouve d'un mois à l'autre les mêmes logements échantillonnés. On utilise l'interview assistée par ordinateur pour recueillir les données d'enquête sur chaque membre des ménages sélectionnés. Ce mode de collecte permet d'obtenir une grande quantité de renseignements sur le PCD auprès des ménages répondants et non répondants.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

à la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

L'EPA est une enquête mensuelle menée selon un plan d'échantillonnage stratifié à plusieurs degrés (Gambino, Singh, Dufour, Kennedy et Lindeyer 1998). Les renseignements utilisés pour construire le plan de sondage et pour prélever un échantillon de logements sont essentiellement géographiques. L'échantillon est divisé en six groupes de renouvellement représentatifs et chaque logement échantillonné reste dans l'échantillon pendant six mois consécutifs. Un groupe de renouvellement contient les logements dont les membres sont interviewés pour la première fois; un autre, ceux dont les membres sont interviewés pour la deuxième fois, et ainsi de suite. C'est pourquoi, dans cinq groupes de renouvellement sur six, on trouve d'un mois à l'autre les mêmes logements échantillonnés. On utilise l'interview assistée par ordinateur pour recueillir les données d'enquête sur chaque membre des ménages sélectionnés. Ce mode de collecte permet d'obtenir une grande quantité de renseignements sur le PCD auprès des ménages répondants et non répondants.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

à la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

à la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

à la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

à la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s, j, \mathbf{D}, \mathbf{Z}, \mathbf{g})$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\mathbf{u}) = [1 + \exp(-\mathbf{u}'(\mathbf{z}_k))]^{-1}$ et l'on suppose que \mathbf{z}_k contient les variables du PCD : le nombre \mathbf{z} des autres. Le vecteur \mathbf{z} contient les variables du PCD des autres. Les variables du plan de sondage fixes \mathbf{d} ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel \mathbf{x} de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus

Il est à noter que l'estimateur aux poids ajustés pour la non-réponse (2.2) est défini implicitement par l'équation

$$U_2(\mathbf{a}, t_{NWVA}^y, t_{NWVA}^x) = t_{NWVA}^y - \sum_{k \in s, y^k} \frac{d^k}{w_k} y^k = 0. \quad (2.3)$$

Si le modèle de non-réponse est spécifié correctement et surtout, si l'hypothèse (2.1) est satisfait, la fonction d'estimation $U_2(\dots)$ est alors sans biais par rapport à $p\#q$ pour t_y ; ainsi, $E^{p\#q}\{U_2(\mathbf{a}, t_y)\} = 0$. Pour rendre l'hypothèse (2.1) aussi plausible que possible, il importe que le modèle de non-réponse soit conditionnel aux variables du plan de sondage, aux variables auxiliaires et aux variables du PCD, aux variables des renseignements déjà contenus dans \mathbf{d} et \mathbf{x} , qui sont bien corrélées avec y , pourvu que ces variables soient également associées à la non-réponse. Cette recommandation devrait être utile pour contrôler l'ampleur du biais de non-réponse, qui peut être inévitable dans une enquête réelle. Elle est aussi compatible avec la recommandation formulée par Little et Vartivarian (2005). Donc, si les variables du PCD contiennent des renseignements sur y au-delà des renseignements déjà contenus dans \mathbf{d} et \mathbf{x} , l'utilisation des variables du PCD peut alors s'avérer utile pour réduire le biais de non-réponse si ces variables sont associées à la non-réponse.

Posons maintenant $\theta = (\mathbf{a}, t_y)$, $\theta = (\mathbf{a}, t_{NWVA}^y)$ et $\bar{U}(\theta) = \{U_1^j(\mathbf{a}, U_2^j(\mathbf{a}, t_y^j))\}$, pour un certain vecteur $\theta = (\mathbf{a}, t_y^j)$. Comme nous l'avons mentionné plus haut, θ est défini implicitement par l'équation $\bar{U}(\theta) = 0$ et la fonction d'estimation $\bar{U}(\cdot)$ est sans biais par rapport à $p\#q$ pour θ puisque $E^{p\#q}\{\bar{U}(\theta)\} = 0$. En utilisant une approximation de Taylor du premier degré (voir Binder 1983), nous avons $\bar{\theta} \approx 0 - \{(\bar{H}(\theta))\}^{-1} \bar{U}(\theta)$, où $\bar{H}(\theta) = E^{p\#q}\{\partial \bar{U}(\theta) / \partial \theta'\}$. La matrice $\{H(\theta)\}^{-1}$ est donc donnée par

$$\{H(\theta)\}^{-1} = \begin{pmatrix} \{H_{11}(\theta)\}^{-1} & -H_{21}(\theta)\{H_{11}(\theta)\}^{-1} \\ 0 & 1 \end{pmatrix}, \quad (2.4)$$

où $H_{11}(\theta) = E^{p\#q}(\partial U_1(\theta) / (\partial \mathbf{a}'))$, pour $i = 1, 2$. En utilisant des conditions semblables à celles de Binder (1983), $\bar{\theta}$ est asymptotiquement normal et asymptotiquement sans biais par rapport à $p\#q$ pour θ . Par conséquent, \bar{y}_{NWVA}^y est asymptotiquement normal et asymptotiquement sans biais par rapport à $p\#q$ pour t_y . Donc, l'utilisation des variables du PCD dans le modèle de non-réponse n'introduit aucun biais dans l'estimateur aux poids ajustés pour la non-réponse \bar{y}_{NWVA}^y pourvu que le modèle de non-réponse (spécification de $q(s, \mathbf{d}, \mathbf{x}, \mathbf{Z}, \mathbf{Z}_s)$ et l'hypothèse 2.1) soit valable. De plus, si le vrai mécanisme de non-réponse inconnu dépend de la partie correspondant à l'échantillon de \mathbf{Y}, \mathbf{Y}_s après avoir conditionné sur s, \mathbf{D}_s et \mathbf{X}_s , le conditionnement sur un vecteur \mathbf{z} de variables du PCD aura tendance à réduire le biais de non-réponse si le

mécanisme du PCD dépend de \mathbf{Y}_s après avoir conditionné sur s, \mathbf{D}_s et \mathbf{X}_s , ce qui signifie que les variables du PCD contiennent des renseignements sur y qui ne sont pas déjà contenus dans \mathbf{d} et \mathbf{x} .

En poursuivant notre linéarisation de Taylor et en utilisant le fait que

$$\begin{aligned} \mathbf{V}^{p\#q}(\bar{U}(\theta)) &= \mathbf{V}^p E^{p\#q}(\bar{U}(\theta) | s) \\ &+ \mathbf{E}^p \mathbf{V}^q E^{p\#q}(\bar{U}(\theta) | s, \mathbf{Z}_s) \\ &+ \mathbf{E}^p \mathbf{V}^q \mathbf{V}^q E^{p\#q}(\bar{U}(\theta) | s, \mathbf{Z}_s), \end{aligned}$$

la matrice de variances-covariances par rapport à $p\#q$ de $\bar{\theta}, \mathbf{V}^{p\#q}(\bar{\theta})$, est approximée par

$$\begin{aligned} \mathbf{V}^{p\#q}(\bar{\theta}) &= \{H(\theta)\}^{-1} \mathbf{V}^p E^{p\#q}(\bar{U}(\theta) | s) \{H'(\theta)\}^{-1} \\ &+ \{H(\theta)\}^{-1} \mathbf{E}^p \mathbf{V}^q E^{p\#q}(\bar{U}(\theta) | s, \mathbf{Z}_s) \{H'(\theta)\}^{-1} \\ &+ \{H(\theta)\}^{-1} \mathbf{E}^p \mathbf{V}^q \mathbf{V}^q E^{p\#q}(\bar{U}(\theta) | s, \mathbf{Z}_s) \{H'(\theta)\}^{-1}. \end{aligned} \quad (2.5)$$

Le premier terme du membre droit de l'équation (2.5) est appelé la variance d'échantillonnage de $\bar{\theta}$, le deuxième, la variance du PCD de $\bar{\theta}$ et le troisième, la variance due à la non-réponse de $\bar{\theta}$. La variance $\mathbf{V}^{p\#q}(t_{NWVA}^y)$ est approximée par la valeur de la dernière ligne et de la dernière colonne de l'équation (2.5). En utilisant l'expression (2.4) et le fait que $\mathbf{E}^q(\bar{U}(\theta) | s, \mathbf{Z}_s) = (\mathbf{0}, t_y - t_y^j)$, la variance approximative (2.5) se réduit à

$$\mathbf{V}^{p\#q}(\bar{\theta}) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^d(t_y^j) \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \mathbf{E}^p \mathbf{V}^q \mathbf{V}^q E^{p\#q}(\bar{U}(\theta) | s, \mathbf{Z}_s) \{H'(\theta)\}^{-1}. \quad (2.6)$$

La deuxième matrice du membre droit de l'équation (2.6) correspond à la variance du PCD de $\bar{\theta}$ et contient 0 pour tous ses éléments. Donc, l'utilisation des variables aléatoires auxiliaires (du PCD) dans le modèle de non-réponse n'introduit aucun terme additionnel de variance, contrairement à l'utilisation de variables auxiliaires fixes uniquement, lorsque le modèle de non-réponse est bien spécifié. Comme les variables du PCD ont tendance à réduire le biais de non-réponse si elles sont associées à y , il semble alors avantageux d'en profiter lorsqu'on traite la non-réponse totale au moyen d'un ajustement de poids. De plus, comme l'ont souligné Little et Vartivarian (2005), l'ajout, dans le modèle de non-réponse, de variables auxiliaires associées à y a tendance à réduire la variance due à la non-réponse. On peut donc réduire l'erreur quadratique moyenne sur les deux fronts.

On peut obtenir une expression plus détaillée du terme de la non-réponse de la variance due à l'échantillonnage et à qu'un estimateur de la variance manie que dans Beaumont (2005). Beaumont (2005) aborde également l'effet de

estime des totaux de population. Cette question du caractère aléatoire des variables auxiliaires du PCD a été soulevée et débattue au sein du Comité consultatif sur les méthodes statistiques de Statistique Canada à la suite de la présentation de l'article d'Alavi et Beaumont (2004). L'objet de la section 2 consiste donc à éclaircir cette question. L'utilisation des variables du PCD pour ajuster les poids de sondage pour la non-réponse est illustrée brièvement dans la section 3, au moyen de l'Enquête sur la population active (EPA) du Canada. La dernière section, soit la section 4, présente un bref résumé de l'article.

2. Théorie

Supposons que nous voulions estimer le total de population $t_y = \sum_{k \in U} y_k$, d'une variable d'intérêt y pour une certaine population fixe U de taille N . De cette population, on sélectionne un échantillon aléatoire s de taille n selon un plan d'échantillonnage probabiliste $p(s|D)$, où D est une matrice à N lignes contenant d^k dans sa k^e ligne et d est le vecteur des variables du plan de sondage. Supposons également qu'en l'absence de non-réponse, nous utiliserions l'estimateur de Horvitz-Thompson $t_y = \sum_{k \in s} w_k y_k$, où $w_k = 1/\pi_k$ est le poids de sondage de l'unité k et $\pi_k = P(k \in s)$ est sa probabilité de sélection.

Habituellement, pour un certain nombre de raisons, la non-réponse totale survient de sorte qu'on observe la variable y uniquement pour un sous-ensemble s_r de s , c'est-à-dire les répondants. Outre s_r , on observe également un vecteur aléatoire z de variables du PCD pour chaque unité de l'échantillon, selon un mécanisme conjoint $\#q(Z_s, s_r | s, Y, D, X)$. Comme nous l'avons mentionné dans l'introduction, le nombre d'essais effectués pour joindre une unité de l'échantillon constitue un exemple de variable du PCD. Le vecteur z des variables du PCD et l'ensemble de répondants s_r sont aléatoires après avoir conditionné sur l'échantillon sélectionné puisque ces quantités prendraient probablement des valeurs différentes si le processus de collecte des données était répété pour un échantillon donné. La quantité Z_s est une matrice à n lignes contenant z^k dans sa k^e ligne, Y est un vecteur à N éléments contenant y_k dans son k^e élément et X est une matrice à N lignes contenant x^k dans sa k^e ligne. Le vecteur x est un vecteur de variables auxiliaires fixes additionnelles. Par exemple, ces variables auxiliaires pourraient provenir d'un fichier administratif ou, dans le cas d'une enquête longitudinale, il pourrait s'agir des variables d'intérêt observées au moment de la vague précédente. Par conséquent, on ne dispose pas nécessairement du vecteur x pour les unités non échantillonnées. Le tableau 1 résume la disponibilité des différents types de variables pour les répondants, les non-répondants et les unités non échantillonnées.

Tableau 1
Disponibilité des variables

	y	z	x	d
Répondants : s_r	OUI	OUI	OUI	OUI
Non-répondants : $s - s_r$	OUI	NON	OUI	OUI
Unités non échantillonnées	NON	NON	NON	OUI**
échantillonnées : $U - s$	NON	NON	NON	OUI

* Le vecteur z n'est même pas défini pour les unités non échantillonnées.
** Le vecteur x peut ne pas toujours être disponible pour les unités non échantillonnées.

On peut factoriser le mécanisme conjoint $\#q(Z_s, s_r | s, Y, D, X)$ en deux mécanismes aléatoires distincts : i) $\#(Z_s | s, Y, D, X)$ et ii) $q(s_r | s, Y, D, X, Z_s)$. Le premier est appelé mécanisme du PCD et le deuxième, mécanisme de non-réponse. Cette factorisation nous permettra plus tard d'obtenir les propriétés de notre estimateur aux poids ajustés pour la non-réponse, défini dans l'équation (2.2) ci-dessous. Nous supposons que

$$t_{NWA}^{t_y} = \sum_{k \in s_r} \frac{d^k(\hat{a})}{w_k} y_k, \quad (2.2)$$

L'estimateur aux poids ajustés pour la non-réponse

Pour compenser la non-réponse totale, nous considérons avoir conditionné sur s_r, D et X .

de sorte qu'il pourrait bien dépendre de Y , même après hypothèse simplificatrice au sujet du mécanisme du PCD, hasard. Toutefois, nous ne formulons explicitement aucune s, D, s_r, X et Z_s , et que les données sont manquantes au Y (ou non confondu avec Y) après avoir conditionné sur plique que le mécanisme de non-réponse est indépendant de dant à l'échantillon de D et de X . Cette hypothèse im- où D et X sont, respectivement, les parties correspon- lité de réponse conditionnelle pour une unité $k \in s$ et \hat{a} est un estimateur du vecteur des paramètres inconnus du mo- dèle de non-réponse a . Il est à noter qu'un modèle de non-réponse est un ensemble d'hypothèses relatives au méca- nisme de non-réponse inconnu $q(s_r | s, Y, D, X, Z_s)$. L'une d'elles est l'hypothèse (2.1). Nous supposons que \hat{a} $U_1(\cdot)$ est un vecteur de fonctions d'estimation sans biais par rapport à q pour a ; ainsi, $E_q\{U_1(a) | s, Y, D, X, Z_s\} = 0$. Donc, $U_1(\cdot)$ est également sans biais par rapport à $p \neq q$ pour a . Dans le reste de l'article, nous supposons partout le conditionnement sur Y, D et X quand on prend les espérances et les variances, puisque ces vecteurs sont toujours considérés comme fixes. Par exemple, nous écrivons $E_q\{U_1(a) | s, Z_s\} = 0$ au lieu de $E_q\{U_1(a) | s, Y, D, X, Z_s\} = 0$. Ceci simplifie considé- rablement la notation.

L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids

Jean-François Beaumont¹

Résumé

On utilise couramment l'ajustement de poids pour la non-réponse afin de compenser la non-réponse totale aux enquêtes. Souvent, on postule un modèle de non-réponse et on ajuste les poids de sondage par l'inverse de probabilités de réponse estimées. Le modèle de non-réponse est habituellement conditionnel à un vecteur de variables auxiliaires fixes qui sont observées pour chaque unité de l'échantillon, comme les variables utilisées pour construire le plan de sondage. Dans le présent article, nous envisageons d'utiliser comme variables auxiliaires éventuelles les variables du processus de collecte des données. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Dans notre traitement, ces variables auxiliaires sont considérées comme aléatoires, même si on conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné. Nous montrons que ce caractère aléatoire n'introduit ni biais ni composante supplémentaire de variance dans les estimations des données. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Dans notre traitement, ces variables auxiliaires sont considérées comme aléatoires, même si on conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné. Nous montrons que ce caractère aléatoire n'introduit ni biais ni composante supplémentaire de variance dans les estimations des données. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Dans notre traitement, ces variables auxiliaires sont considérées comme aléatoires, même si on conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné. Nous montrons que ce caractère aléatoire n'introduit ni biais ni composante supplémentaire de variance dans les estimations des données. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Dans notre traitement, ces variables auxiliaires sont considérées comme aléatoires, même si on conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné.

Mots clés : Biais de non-réponse; modèle de non-réponse; variance due à la non-réponse; nombre d'essais; paramètres; probabilité de réponse.

1. Introduction

Dans les enquêtes, on traite souvent la non-réponse totale en utilisant une méthode d'ajustement de poids pour la non-réponse. Le principe de base qu'on choisit souvent consiste à ajuster les poids de sondage par l'inverse de probabilités de réponse estimées (voir, par exemple, Ekholm et Laaksonen 1991). On obtient ces probabilités de réponse estimées en postulant un modèle pour le mécanisme de non-réponse inconnu, que nous appelons le modèle de non-réponse. Pour réduire dans toute la mesure du possible le biais et la variance dus à la non-réponse, il est essentiel de conditionner sur un vecteur de variables auxiliaires qui sont observées pour chaque unité de l'échantillon et qui sont de bons prédicteurs de la non-réponse et des variables d'intérêt (Little et Vartavarian 2005). On traite habituellement les variables auxiliaires comme des variables fixes, que ce soit conditionnellement ou non à l'échantillon sélectionné.

Dans le présent article, nous envisageons d'utiliser les variables du processus de collecte des données (PCD) comme variables auxiliaires éventuelles à inclure dans le modèle de non-réponse. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Ce type de données est parfois appelé paradoxes (voir Couper et Lyberg 2005 pour une référence récente sur le sujet); Holt et Elliott (1991), entre autres, l'ont utilisé pour

composer avec la non-réponse totale. Dans notre traitement, contrairement à Holt et Elliott (1991), les variables du PCD sont considérées comme aléatoires, même si on les conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné.

Les variables du PCD peuvent s'avérer particulièrement utiles dans les enquêtes transversales où les variables auxiliaires dont on dispose pour traiter la non-réponse totale se limitent souvent aux variables utilisées pour construire le plan de sondage. Sans être inutiles, ces variables du plan de sondage ne sont souvent pas de bons prédicteurs de la non-réponse et des variables d'intérêt. Dans ce cas, les renseignements supplémentaires tirés du processus de collecte des données peuvent être les bienvenus. Dans les enquêtes longitudinales, on trouve une foule de variables auxiliaires éventuelles pour composer avec la non-réponse de vague. Les renseignements sur le PCD peuvent donc s'avérer moins importants pour compenser la non-réponse de vague que pour compenser la non-réponse totale dans les enquêtes transversales, mais nous n'avons pas encore étudié cet aspect en profondeur. Il se pourrait qu'aux points de changement, les variables du PCD jouent un rôle important.

Dans la section 2, nous présentons la notation et notre théorie concernant l'effet de l'utilisation de variables auxiliaires aléatoires dans le modèle de non-réponse lorsqu'on

- Farahmand, B.Y., Persson, P.G., Michaëlsson, K., Baron, J.A., Parker, M.G. et Ljunghall, S. (2000). Socioeconomic status, marital status and hip fracture risk: A population-based case control study. *Osteoporosis International*, 11, 803-808.
- Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society, Series B*, 60, 57-70.
- Garny, O., Baudoin, C. et Fardellone, P. (2000). Effect of alcohol intake on bone mineral density in elderly women: The EPIDOS Study. *American Journal of Epidemiology*, 151, 8, 773-780.
- Kass, R., et Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Kallion, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.
- Lauderdale, D.S., et Rathouz, P.J. (2003). Does bone mineralization reflect economic conditions? An examination using a national US sample. *Economics and Human Biology*, 1, 91-104.
- Little, R.J.A., et Rubin D.B. (2002). *Statistical Analysis with Missing Data*. Édition, New York: John Wiley & Sons, Inc.
- Little, R.J.A., et Rubin D.B. (2002). *Statistical Analysis with Missing Data*. Édition, New York: John Wiley & Sons, Inc., 289-306.
- Dans *Analysis of Survey Data*, (Eds. R.L. Chambers et C.J. Skinner), New York: John Wiley & Sons, Inc., 1761-1768.
- Looker, A.C., Wahner, H.W., Dunn, W.L., Calvo, M.S., Harris, R.R., Heyse, S.P., Johnston, C.C. et Lindsay, R. (1998). Updated data on proximal femur bone mineral levels of us adults. *Osteoporosis International*, 8, 468-489.
- Mitkin, B. (2001). Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, 55, 111-120.
- Nandram, B., et Choi, J.W. (2002 a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Wang, H. (2001). Two-way Contingency Tables with Marginal and Conditionally Imputed Nonrespondents. Thèse de doctorat, Department of Statistics, University of Wisconsin-Madison.
- Sinharay, S., et Stern, H.S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56, 196-201.
- Rubin, D.B., Stern, H.S. et Vehovar, V. (1995). Handling "Don't know" survey responses: The case of the slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Ritter, C., et Tanner, M.A. (1992). The Gibbs stopper and the gridly Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.
- Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multivariate contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Nandram, B., Liu, N., Choi, J.W. et Cox, L.H. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.
- Nandram, B., et Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, 72, 319-340.
- Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayésien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.
- Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayésien hiérarchique de non-réponse non-ignorable pour les petites domaines : Une application aux données de la NHANES. *Techniques d'enquête*, 31, 79-92.
- Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour les proportions under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.
- Nandram, B., et Choi, J.W. (2002 b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

$\phi/(1-\phi)$, les paramètres résident sur $(0,1)$ avec les contraintes appropriées, ce qui rend la méthode du quadrillage commode. Nous utilisons 50 intervalles de même largeur (obtenus par expérimentation) pour tirer μ et ϕ , et une valeur aléatoire pour τ est $\phi/(1-\phi)$. Nous exécutons l'échantillonnage de Gibbs en tirant une valeur aléatoire de chacune des « densités » à posteriori conditionnelles, (A.1), (A.2), (A.3) et (A.4) l'une après l'autre, et en itérant la procédure complète jusqu'à la convergence. Il s'agit d'un exemple d'échantillonnage « gnddy Gibbs » (Ritter et Tanner 1992).

Annexe B

Estimation de $p_{\text{NIC}}(y_1)$ dans (16)

En notant n_m le nombre de cas incomplets (c'est-à-dire $n = n_0 + n_m$), nous pouvons aussi montrer que, pour le modèle avec association, $p_{\text{NIC}}(y_1) = a/(n+1)/(n_0!n_m!)$; A et pour le modèle sans association, $p_{\text{NIC}}(y_1) = b/(n+1)!/(n_0!n_m!)$; B , où a et b sont donnés par (18),

$$A = \int \omega_a \left\{ \prod_{j,k} \pi_{y_{j,k}}^{j,k} \right\} \left\{ \sum_{s=2}^4 \sum_{j,k} \pi_{y_{j,k}}^{s,j,k} p_{j,k} \right\} \left\{ \frac{D(y_{111} + 1, \dots, y_{1rc} + 1)}{\prod_{j,k} p_{y_{j,k}}^{j,k}} \right\} \times \prod_{j,k} \left\{ \frac{D(\mu \tau)}{\prod_{s,j,k} \pi_{y_{j,k}}^{s,j,k}} \right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0-1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_a,$$

$$B = \int \omega_{na} \left\{ \prod_{j,k} \pi_{y_{j,k}}^{j,k} \right\} \left\{ \sum_{s=2}^4 \sum_{j,k} \pi_{y_{j,k}}^{s,j,k} q_{j,k}^{s,j,k} \right\} \left\{ \frac{D(y_{11} + 1, \dots, y_{1,c} + 1)}{\prod_{j,k} q_{y_{j,k}}^{j,k}} \right\} \times \frac{D(y_{11} + 1, \dots, y_{1,c} + 1)}{\prod_{j,k} q_{y_{j,k}}^{j,k}} \left\{ \frac{D(\mu \tau)}{\prod_{s,j,k} \pi_{y_{j,k}}^{s,j,k}} \right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0-1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_{na}.$$

Notons que $0 < A, B < 1$ donne une vérification diagnostique utile des calculs. Nous montrons comment calculer A dans (B.1) par la méthode d'intégration de Monte Carlo; la méthode pour calculer B est semblable. Nous préférons la méthode simple fondée sur l'intégration par la méthode de Monte Carlo avec une fonction d'importance (Nandram et Kim, 2002) à celle fondée sur une continuation de l'échantillonneur de Gibbs (Chib et Jeliazkov 2001).

Remerciements

La matière présentée ici fait partie des travaux réalisés au cours de l'année universitaire 2003-2004 durant laquelle Balgobin Nandram était en congé sabbatique à titre de chercheur au National Center for Health Statistics, à Hyattsville, et les deux examinateurs de leurs commentaires adjoint et les deux examinateurs de leurs commentaires constructifs et des trois occasions que nous avons eues de réviser le manuscrit.

Bibliographie

Chen, T., et Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Chib, S., et Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96, 270-281.

Cohen, G., et Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents. *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45-97.

information a priori) au sujet de la non-ignorabilité, des poids de sondage et des effets de mise en grappes également.

Annexe A Ajustement du modèle de non-réponse non ignorable

Nous montrons comment utiliser l'échantillonneur de Gibbs pour faire une inférence au sujet des paramètres de (14). La densité a posteriori conditionnelle de p est

$$p \mid y \sim \text{Dirichlet}(y_{11} + 1, \dots, y_{rc} + 1) \quad (\text{A.1})$$

et la densité a posteriori conditionnelle de π_{jk} est

$$\pi_{jk} \mid \mu, \tau, y \sim \text{Dirichlet} \left(y_{1jk} + \mu_{1\tau}, y_{2jk} + \mu_{2\tau}, y_{3jk} + \mu_{3\tau}, y_{4jk} + \mu_{4\tau} \right) \quad (\text{A.2})$$

avec indépendance sur $j = 1, \dots, r, k = 1, \dots, c$.

Nous avons besoin des fonctions de masse de probabilité a posteriori conditionnelles de $y_s, s = 2, 3, 4$ sachant $y_{(s)}, \mathbf{d}, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c$. D'après (14), il est clair que les $y_s, s = 2, 3, 4$ sont des vecteurs aléatoires multinomiaux conditionnellement indépendants. Plus précisément,

$$y_{2j} \mid \{y_1, \mathbf{d}, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c\}$$

$$\stackrel{\text{ind}}{\sim} \text{Multinomial}(n_j, \mathbf{q}_{(c)}^j), j = 1, \dots, r,$$

$$y_{3k} \mid \{y_1, \mathbf{d}, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c\}$$

$$\stackrel{\text{ind}}{\sim} \text{Multinomial}(v_k, \mathbf{q}_{(3)}^k), k = 1, \dots, c,$$

$$y_4 \mid \{y_1, \mathbf{d}, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c\}$$

$$\sim \text{Multinomial}(w, \mathbf{g}^{(4)}), \quad (\text{A.3})$$

$$\text{où } \mathbf{g}^{(2)} = \pi_{2jk} / \sum_{k=1}^c \pi_{2jk}, \mathbf{g}^{(3)} = \pi_{3jk} / \sum_{j=1}^r \pi_{3jk}, \mathbf{g}^{(4)} = \pi_{4jk} / \sum_{j=1}^r \sum_{k=1}^c \pi_{4jk} \text{ et } \mathbf{q}_{(4)}^k = \pi_{4jk} / \sum_{j=1}^r \pi_{4jk}, j = 1, \dots, r, k = 1, \dots, c.$$

Puis, nous considérons les hyperparamètres. Si nous posons que $\delta_s = \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}$, la densité a posteriori conditionnelle conjointe de μ, τ est

$$p(\mu, \tau \mid \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c) \propto \left[\prod_{j=1}^r \delta_{\mu, s}^j / D(\mu, \tau) \right] \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}, \quad (\text{A.4})$$

$$\text{où } \sum_{s=1}^4 \mu_s = 1, \mu_s \geq 0, s = 1, 2, 3, 4, \tau > 0.$$

Nous utilisons la méthode du quadrillage pour obtenir des échantillons à partir de la densité a posteriori conditionnelle de $p(\mu \mid \tau, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c)$ et $p(\tau \mid \mu, \pi_{jk}, j = 1, \dots, r, k = 1, \dots, c)$. Après transformation de τ en

variables DMO et RF. Pour l'ensemble des données, notre facteur de Bayes indique que la preuve d'absence d'association est « forte » sous le modèle de non-réponse ignorable et qu'elle est « positive » sous le modèle de non-réponse non ignorable. Donc, il n'y a pour ainsi dire aucune différence entre les deux scénarios, c'est-à-dire celui où seules les données provenant des cas complets sont utilisées et celui où toutes les données sont utilisées. D'après le facteur de Bayes et notre étude par simulation, bien qu'il existe des différences entre le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable, elles sont faibles. Nous constatons des différences en ce qui concerne l'inférence au sujet des proportions d'individus à divers niveaux DMO-RF; les moyennes a posteriori sont semblables, mais les écarts-types a posteriori sont plus grands sous le modèle de non-réponse non ignorable que sous le modèle de non-réponse ignorable.

Notre étude par simulation confirme deux propriétés (différences subtiles) de nos modèles. En premier lieu, les estimations des probabilités de cellule d'après le modèle de non-réponse ignorable (non ignorable) s'approchent plus des valeurs réelles quand il est attendu que le modèle de non-réponse ignorable (non ignorable) sera vérifié, mais, dans l'un et l'autre cas, l'écart-type des estimations d'après le modèle de non-réponse non ignorable est environ deux fois plus grand que celui des estimations d'après le modèle de non-réponse ignorable. En deuxième lieu, si l'on s'attend à ce que le modèle de non-réponse ignorable (non ignorable) tienne, celui-ci peut donner de moins bons résultats que le modèle de non-réponse non ignorable (ignorable). Cela se produit un pourcentage significativement plus élevé de fois dans le cas où l'on s'attend à ce que le modèle de non-réponse ignorable soit vérifié que dans l'autre. Donc, il existe des différences entre ces modèles. Nous suggérons d'ajuster les deux modèles et de calculer le facteur de Bayes pour décider lequel il convient d'utiliser. Nous ne recommandons pas d'utiliser ces modèles lorsqu'il existe des covariables et (ou) des données a priori appropriées pour expliquer la non-ignorabilité.

Lors de futurs travaux, nous pourrions essayer de réduire le nombre de paramètres du modèle de non-réponse non ignorable afin de réduire davantage les effets de la non-ignorabilité. Ainsi, nous pourrions envisager de représenter les données dans deux tableaux de contingence comme il suit. Les trois tableaux supplémentaires sont regroupés en un tableau supplémentaire unique dont la j^{e} ligne contient n_j individus et la k^{e} colonne, au moins v_k individus; le nombre total d'individus dans ce tableau supplémentaire est $w + \sum_{j=1}^r n_j + \sum_{k=1}^c v_k$; voir la section 3.1 pour la notation. Enfin, il convient de souligner que l'analyse complète des données provenant d'une enquête complexe nécessite une apport d'information (covariables et

Tableau 7
Comparaison des modèles de non-réponse ignorable et non ignorable au moyen de données simulées et des moyennes a postérieur (MP) ainsi que des écarts-types a postérieur (ETP) des p^jk

Cellule	Simulée		Ignorable (a)		Non ignorable		Ignorable		Non ignorable	
	Ajusté	\hat{p}	ETP	MP	ETP	MP	ETP	MP	ETP	MP
(1, 1)	321,81	320,73	5,72	307,42	11,30	332,02	5,10	324,44	10,60	324,44
(1, 2)	142,66	142,96	4,24	146,44	7,34	141,81	3,30	143,44	5,43	143,44
(1, 3)	173,40	172,59	4,42	173,49	7,62	168,66	4,14	174,10	7,04	174,10
(2, 1)	138,57	138,82	4,81	135,32	9,82	143,63	4,52	139,20	9,74	139,20
(2, 2)	68,44	68,44	3,55	72,01	6,02	64,51	2,91	68,20	4,76	68,20
(2, 3)	71,11	71,11	3,65	75,00	6,30	70,85	3,76	69,63	6,58	69,63
(3, 1)	52,14	52,17	3,11	53,03	4,95	53,08	3,04	52,44	4,70	52,44
(3, 2)	18,96	19,35	2,08	21,65	2,98	15,08	1,72	17,32	2,48	17,32
(3, 3)	12,93	13,54	1,78	15,64	2,55	10,95	1,85	11,20	2,18	11,20

Note : Les données sont simulées à partir du modèle de non-réponse ignorable en (a) ou du modèle de non-réponse non ignorable en (b), et les modèles de non-réponse ignorable et non ignorable sont tous deux ajustés. Nous avons généré 1 000 ensembles de données et nous avons ajusté le modèle de non-réponse ignorable ainsi que le modèle de non-réponse non ignorable à chaque ensemble de données simulé. Les MP et les ETP sont les moyennes sur les 1 000 ensembles de données et \hat{p} est la moyenne a postérieur pour les données observées que nous avons utilisées pour générer les ensembles de données. Toutes les entrées doivent être multipliées par 10³.

ignorable (non ignorable) donne une MP de 0,321 (0,307), mais dans (b), le modèle de non-réponse ignorable (non ignorable) donne une MP de 0,332 (0,324) pour d'autres exemples. Donc, les deux modèles donnent effectivement des résultats différents lors de l'estimation de p .

Nous avons également considéré l'estimation de la proportion P d'ensembles de données simulés dans lesquels le modèle de non-réponse ignorable donne de meilleurs résultats que le modèle de non-réponse non ignorable. Il est coûteux de calculer la vraisemblance marginale sous le modèle de non-réponse non ignorable. Nous souignons de nouveau qu'il faut 50 000 itérations pour que l'estimation de Monte Carlo se stabilise; il s'agit là d'une tâche énorme pour l'étude par simulation, parce que nous devons calculer les vraisemblances marginales pour 1 000 ensembles de données. Donc, nous utilisons une méthode simple pour comparer les deux modèles et nous nous attendons à ce qu'elle donne une conclusion comparable à un calcul

puissant.

Plus précisément, nous calculons $\Delta_{(h)} = \sum_{j=1}^J \sum_{k=1}^K (p^{jk} - PM_{(h)}^{jk})^2 / PM_{(h)}^{jk}$, où $PM_{(h)}^{jk}$ est la moyenne a postérieur de p^{jk} correspondant au h^e ensemble de données. Nous notons $\Delta_{(h)}^{IG}$ par $\Delta_{(h)}^{IG}$ pour le modèle de non-réponse ignorable et par $\Delta_{(h)}^{NIG}$ pour le modèle de non-réponse non ignorable. Nous obtenons un estimateur de P , \hat{p} , en comptant le nombre d'expériences parmi les 1 000 réalisées pour lesquelles $\Delta_{(h)}^{IG} > \Delta_{(h)}^{NIG}$. Pour les données

répondre ignorable soit vérifié, il sera battu par le modèle de non-réponse ignorable à ce que le modèle de non-réponse ignorable, \hat{p} est égal à 0,920 avec une erreur-type de 0,009. Donc, si nous nous attendons à ce que le modèle de non-réponse ignorable soit vérifié, il sera battu par le modèle de non-réponse ignorable à ce que le modèle de non-réponse non ignorable, \hat{p} est égal à 0,236 avec une erreur-type de 0,013. Pour les données générées à partir du modèle de non-réponse non ignorable, \hat{p} est égal à 0,920 avec une erreur-type de 0,009.

non-réponse non ignorable environ 24 % du temps, et si nous nous attendons à ce que le modèle de non-réponse non ignorable soit vérifié, il ne sera battu par le modèle de non-réponse ignorable qu'environ (1-0,920) 100 %, soit environ 8 % du temps. Par conséquent, il existe des différences latentes entre les deux modèles. Le modèle de non-réponse non ignorable reflète un certain degré de non-ignorable et rend le modèle de non-réponse ignorable plus robuste. Nous considérons qu'il s'agit d'une comparaison raisonnable entre les deux modèles.

5. Conclusion

Deux nouvelles méthodes méthodologiques importantes sont exposées dans le présent article. Plus précisément, nous avons montré a) qu'il est possible d'analyser des données multinomiales provenant de tableaux de contingence $r \times c$ en présence à la fois de non-réponse partielle et totale et que le mécanisme de non-réponse peut être non ignorable, et b) qu'en utilisant le facteur de Bayes (ratio des vraisemblances marginales des deux modèles), nous pouvons vérifier s'il existe une association entre les deux caractéristiques. Essentiellement, nous avons supposé qu'il n'existait aucune information au sujet de la non-ignorable, nous avons supprimé toutes les caractéristiques du plan de sondage et nous avons adopté une approche prudente. Pour le tableau de contingence 3×3 contenant des données catégoriques sur la densité minérale osseuse (DMO) et le revenu familial (RF), nous avons montré comment estimer exactement les probabilités de cellule. Pour les cas de données complètes, le facteur de Bayes donne une « forte » preuve d'absence d'association entre les

Tableau 5
Sensibilité des moyennes a posteriori (MP) et ainsi que des écarts-types a posteriori (ETP) des p_{jk} au choix de k dans le modèle de non-réponse non ignorable

k	0,25	0,50	1,00	2,00	4,00
Cellule	306,93	315,01	321,81	325,37	326,16
(1, 1)	306,93	315,01	321,81	325,37	326,16
(1, 2)	141,12	139,86	142,66	142,63	143,42
(1, 3)	161,68	167,83	173,40	176,20	175,78
(2, 1)	143,18	142,62	138,57	137,23	137,26
(2, 2)	68,46	71,06	68,44	68,79	68,11
(2, 3)	79,78	75,97	71,11	68,09	68,34
(3, 1)	59,97	53,50	52,14	50,97	51,41
(3, 2)	21,43	20,02	18,96	23,28	17,84
(3, 3)	17,45	10,38	4,28	2,99	1,99

Nota: Toutes les entrées doivent être multipliées par 10^{-3} . Dans le modèle de non-réponse non ignorable, $\pi_{sjk} \sim \text{Gamma}(k\alpha_0, \beta_0)$, où k est le paramètre de sensibilité et $\alpha_0 = 1,25$ et $\beta_0 = 0,35$.

Dans l'ensemble, il existe un certain niveau de preuve de l'absence d'association. Donc, il est intéressant de savoir que l'on ne doit pas s'inquiéter trop du choix de (α_0, β_0) .

Tableau 6
Sensibilité des vraisemblances marginales et du facteur de Bayes au choix de k dans le modèle de non-réponse non ignorable

k	Association	Pas d'association	Facteur de Bayes
0,25	-53,37	-49,16	-4,21
0,50	-52,58	-49,49	1,82
1,00	-52,58	-49,76	1,79
2,00	-52,81	-49,83	1,78
4,00	-52,95	-49,91	1,77

Nota: Toutes les entrées sont exprimées sur l'échelle logarithmique. Dans le modèle de non-réponse non ignorable, $\pi_{ijk} \sim \text{Gamma}(k\alpha_0, \beta_0)$, où k est le paramètre de sensibilité et $\alpha_0 = 1,25$ et $\beta_0 = 0,35$.

4.3 Étude par simulation

Nous avons exécuté une étude par simulation pour poursuivre la comparaison entre les modèles de non-réponse ignorable et non ignorable. L'objectif est de confirmer qu'il existe des différences entre les deux modèles. Dans notre situation, un test fondé sur le facteur de Bayes permettra d'indiquer si ces différences existent ou non. Lorsque l'information au sujet de la non-ignorabilité est limitée (ce qui est le cas ici), il est raisonnable d'ajuster un modèle de non-réponse ignorable, parce que les paramètres sont identifiables dans ce modèle. Donc, nous procédons à la comparaison lorsque les données sont générées à partir du modèle de non-réponse ignorable et b) du modèle de non-réponse non ignorable. Il s'agit d'une analyse bayésienne typique. Nous obtenons les moyennes a posteriori des p_{jk} et des π_{sjk} , notées \hat{p}_{jk} et $\hat{\pi}_{sjk}$, respectivement, après avoir ajusté nos modèles de non-réponse non ignorable aux données observées. Pour la non-réponse ignorable, nous prenons $\pi_{js} = \sum_{j=1}^J \pi_{sjk} / r_{rc}$, $s = 1, 2, 3, 4$. Nous obtenons les

effectifs de cellules pour le modèle de non-réponse ignorable à partir de $(Y_{111}, \dots, Y_{1rc}, \dots, Y_{411}, \dots, Y_{4rc}) | \pi, \hat{p}$
 $\sim \text{Multinomial}(n, (\pi_1 \hat{p}_{11}, \dots, \pi_4 \hat{p}_{rc}))$
et pour le modèle de non-réponse non ignorable, par tirage à partir de $(Y_{111}, \dots, Y_{1rc}, \dots, Y_{411}, \dots, Y_{4rc}) | \pi, \hat{p}$
 $\sim \text{Multinomial}(n, (\pi_1 \hat{p}_{11}, \dots, \pi_4 \hat{p}_{rc}))$

~Multinomial $\{n, (\pi_{111} \hat{p}_{11}, \dots, \pi_{4rc} \hat{p}_{rc})\}$, où $n = 2\,998$, le nombre total d'individus dans l'ensemble de données original (voir le tableau 1). Nous avons généré 1 000 ensembles de données pour le modèle de non-réponse ignorable ainsi que pour le modèle de non-réponse non ignorable. Puis, nous avons ajusté les modèles de non-réponse ignorable et non ignorable à chaque ensemble de données exactement de la même manière que pour les données observées du tableau 1 et nous avons calculé les moyennes a posteriori (MP) et les écarts-types a posteriori (ETP) pour les p_{jk} . Au tableau 7, nous présentons les moyennes des MP et des ETP sur les 1 000 ensembles de données. La deuxième colonne (étiquetée \hat{p}) contient la moyenne a posteriori de p_{jk} pour les données observées sous le modèle de non-réponse non ignorable (voir le tableau 2b).

Pour (a) au tableau 7, les MP sont très proches des \hat{p}_{jk} pour le modèle de non-réponse ignorable, mais pas autant si l'on ajuste le modèle de non-réponse non ignorable. Il est évident que les ETP sont environ deux fois plus grands sous le modèle de non-réponse non ignorable que sous le modèle de non-réponse ignorable. Pour (b) au tableau 7, les MP s'approchent plus des \hat{p}_{jk} pour le modèle de non-réponse non ignorable que pour le modèle de non-réponse ignorable. Cependant, dans les deux cas, les ETP sont environ deux fois plus grands pour le modèle de non-réponse non ignorable que pour le modèle de non-réponse ignorable. Par exemple, au tableau 7 pour la cellule (1, 1) comparativement à 0,322 pour \hat{p} , dans (a), le modèle de non-réponse

Kass et Raftery (1995). Donc, il existe de nouveau une différence entre les modèles de non-réponse ignorable et non ignorable. Toutefois, l'ETN de 1,80 a tendance à annuler ces différences. Nous concluons qu'il existe des données convaincantes montrant qu'il n'y a pas d'association entre la densité minérale osseuse (DMO) et le revenu familial (RF).

Tableau 3

Comparaison des moyennes a posteriori (MP) et des écarts-types a posteriori (π_{jk}^{ETP}) tiré des modèles de non-réponse ignorable et non ignorable

Non ignorable	
π_1	0,615 (0,009)
π_2	0,077 (0,005)
π_3	0,292 (0,008)
π_4	0,015 (0,002)
Non ignorable	
π_1	0,388 (0,078)
π_2	0,057 (0,017)
π_3	0,195 (0,068)
π_4	0,217 (0,041)
Non ignorable	
π_1	0,656 (0,044)
π_2	0,057 (0,017)
π_3	0,195 (0,068)
π_4	0,217 (0,041)
Non ignorable	
π_1	0,388 (0,078)
π_2	0,057 (0,017)
π_3	0,195 (0,068)
π_4	0,217 (0,041)
Non ignorable	
π_1	0,656 (0,044)
π_2	0,057 (0,017)
π_3	0,195 (0,068)
π_4	0,217 (0,041)

Nota : Les ETP figurent entre parenthèses. Pour le modèle de non-réponse ignorable, les paramètres sont π_1, π_2, π_3 et π_4 , et pour le modèle de non-réponse non ignorable, les paramètres sont $\pi_1, \pi_2, \pi_3, \pi_4, \pi_5$ et π_6 . Pour chaque s , nous avons sélectionné parmi les neuf cellules la plus petite et la plus grande MP pour former l'intervalle.

Tableau 4

Vraisemblances marginales et facteurs de Bayes pour le test d'association entre DMO et RF sous les modèles de non-réponse ignorable et non ignorable

Association		Pas d'association		Différence	
Ignorable	-49,571	-46,173	-3,398	Non ignorable	-53,129
ETN	1,800	1,790		ETN	1,800

Nota : Toutes les cellules (vraisemblances marginales et leurs différences) sont exprimées sur l'échelle logarithmique. L'intégration par la méthode de Monte Carlo comprend 50 000 itérations. Les ceneurs-types numériques (ETN) sont faibles comparativement aux vraisemblances marginales.

Nous avons examiné la relation entre DMO et RF quand les niveaux d'ostéoporose sont regroupés en un seul. Sous le modèle de non-réponse ignorable, le logarithme du facteur de Bayes est égal à -2,77 (log vraisemblance marginale : -32,82 et -29,05) et sous le modèle de non-réponse non ignorable, il est égal à -4,52 (log vraisemblance marginale : -34,25 et -4,52). Donc, nous arrivons à la même conclusion au sujet de l'absence d'association entre DMO et RF.

Nous avons également réparti les données en deux groupes d'âge, c'est-à-dire les femmes préménopausées (ayant, au plus, 49 ans; jeunes) et les femmes ménopausées (ayant au moins 50 ans; âgées). Parmi le groupe de femmes jeunes, quatre seulement faisaient de l'ostéoporose, si bien que nous avons regroupé celles faisant de l'ostéopénie et celles faisant de l'ostéoporose. Nous avons ajusté le modèle de non-réponse ignorable ainsi que le modèle de non-réponse non ignorable à ces données et obtenu des résultats

comparables. Pour le groupe de femmes âgées, en utilisant le modèle de non-réponse ignorable, les logarithmes des vraisemblances marginales correspondant à l'absence d'association et à l'existence d'une association sont -43,01 et -38,91, ce qui donne un logarithme du facteur de Bayes de 4,10 pour l'absence d'association. Par conséquent, il existe de fortes preuves d'absence d'association entre DMO et RF. Pour le groupe de femmes jeunes, si nous utilisons le modèle de non-réponse ignorable, les logarithmes des vraisemblances marginales correspondant à l'absence d'association et à l'existence d'une association sont -29,93 et -28,80, ce qui donne un logarithme du facteur de Bayes de 1,13 pour l'absence d'association. Donc, il existe des indices positifs d'une absence d'association entre la densité minérale osseuse et le revenu familial pour les deux groupes d'âge. Par conséquent, il est peu probable que l'âge joue un rôle dans l'association entre les deux variables.

4.2 Analyse de sensibilité

Nous avons étudié la sensibilité de l'inférence au sujet de p_{jk} à la loi a priori de τ . Autrement dit, nous avons pris $\tau \sim \text{Gamma}(\kappa\alpha_0, \beta_0)$, où κ est un paramètre de sensibilité auquel nous avons donné la valeur de 1 dans notre analyse (notons que $E(\tau) = \kappa\alpha_0 / \beta_0$).

Notre méthode pour la spécification de α_0 et β_0 donne les valeurs de $\alpha_0 = 125$ et $\beta_0 = 0,35$; voir la section 3.5. Donner à κ une valeur supérieure à 1 induit des variations moins importantes de la moyenne a posteriori (MP) et de l'écart-type a posteriori (ETP) des p_{jk} que lui donner une valeur inférieure à 1, parce que les valeurs plus élevées de κ provoquent des variations beaucoup plus faibles de la loi a priori de τ . Au tableau 5, nous présentons les MP et les ETP des p_{jk} pour $\kappa = 0,25, 0,50, 1,00, 2,00$ et 4,00. La valeur des MP augmente avec κ et celle des ETP diminue lorsque κ passe de 0,25 à 4,00. Donc, il existe une certaine sensibilité à la spécification de α_0 et de β_0 , mais les variations sont faibles. Par exemple, les MP de p_{11} sont 0,31, 0,32 et 0,33 pour $\kappa = 0,25, 1,00$ et 4,00, et, à ces valeurs de κ , les ETP sont 0,04, 0,02 et 0,01.

Nous avons également étudié la sensibilité des facteurs de Bayes au choix de κ (voir le tableau 6). Pour commencer, les ETN diminuent avec κ , mais la variation est faible. Notons que nous avons utilisé 50 000 itérations pour l'intégration par la méthode de Monte Carlo; cette taille d'échantillon est nécessaire pour que les estimations de Monte Carlo se stabilisent. Les logarithmes des vraisemblances marginales varient peu avec κ . Comme les logarithmes des facteurs de Bayes sont faibles, certaines variations sont reflétées dans l'inférence : pour $\kappa = 0,25, 0,50$ et 4,00, il existe de « fortes » preuves de l'absence d'association, mais pour $\kappa = 1,00$ et 2,00, il existe des « positives » preuves de l'absence d'association.

pour l'autre, ce qui rend les intervalles de confiance à 95 % ainsi dire tous les intervalles de confiance à 95 % sont contenus dans ceux obtenus pour le modèle de non-réponse non ignorable.

Tableau 2

Comparaison des moyennes a posteriori (MP), des écarts-types a posteriori (ETP), des erreurs types numériques (ETN) et des intervalles de confiance à 95 % (IC) pour p tiré des modèles de non-réponse ignorable et non ignorable

Cellule	p̂	MP	ETP	ETN	IC
a) Modèle de non-réponse ignorable	0,337	0,330	0,005	0,001	(0,321, 0,339)
(1, 1)	0,157	0,142	0,003	0,001	(0,136, 0,147)
(1, 2)	0,157	0,142	0,003	0,001	(0,162, 0,175)
(2, 1)	0,141	0,142	0,004	0,001	(0,134, 0,148)
(2, 2)	0,071	0,066	0,002	0,001	(0,061, 0,070)
(2, 3)	0,063	0,071	0,003	0,001	(0,066, 0,078)
(3, 1)	0,050	0,053	0,003	0,001	(0,048, 0,059)
(3, 2)	0,016	0,016	0,001	0,000	(0,013, 0,019)
(3, 3)	0,010	0,012	0,002	0,000	(0,009, 0,015)
b) Modèle de non-réponse non ignorable	0,321	0,320	0,009	0,000	(0,278, 0,355)
(1, 1)	0,157	0,143	0,008	0,003	(0,126, 0,158)
(1, 2)	0,154	0,173	0,007	0,004	(0,140, 0,196)
(2, 1)	0,141	0,139	0,019	0,009	(0,109, 0,182)
(2, 2)	0,071	0,069	0,007	0,003	(0,056, 0,085)
(2, 3)	0,063	0,071	0,013	0,006	(0,053, 0,102)
(3, 1)	0,050	0,052	0,008	0,002	(0,040, 0,070)
(3, 2)	0,016	0,019	0,003	0,001	(0,014, 0,026)
(3, 3)	0,010	0,013	0,003	0,001	(0,009, 0,020)

Nota : Pour le modèle de non-réponse ignorable, $\pi_{ijk} = \pi_s$, $s = 1, 2, 3$. La valeur observée de p basée sur les données complètes est p̂.

Au tableau 3, nous comparons également l'estimation de

π_s dans le modèle de non-réponse ignorable à π_{ijk}^{ijk} dans le modèle de non-réponse non ignorable. Pour ce dernier, nous présentons la fourchettes des moyennes a posteriori (MP) pour les neuf cellules de chaque s, s = 1, 2, 3, 4. Elle indique l'importance de la non-ignorabilité. Les MP de π_s sont comprises dans la fourchette des π_{ijk}^{ijk} et, comme prévu, les ETP sont plus grands pour le modèle de non-réponse non ignorable que pour l'autre. Par exemple, sur les neuf cellules, les π_{ijk}^{ijk} varient de 0,388 à 0,656, et ces deux chiffres diffèrent significativement de 0,615, ce qui témoigne d'un certain degré de non-ignorabilité. Donc, il existe une différence entre les modèles de non-réponse ignorable et non ignorable.

Au tableau 4, nous présentons les logarithmes des facteurs de Bayes utilisés pour tester la qualité de l'ajuste-ment du modèle de non-réponse ignorable et du modèle de non-réponse non ignorable. Il existe de « fortes » preuves que le modèle de non-réponse ignorable est mieux ajusté que le modèle de non-réponse non ignorable aux données étudiées (Kass et Raftery 1995). Alors que le modèle de non-réponse ignorable donne de « fortes » preuves d'absence d'association, le modèle de non-réponse non ignorable donne un résultat « positif », comme l'indique

4. Données et analyse empirique

Nous appliquons notre méthode aux données du tableau de contingence 3×3 illustré au tableau 1. Après avoir présenté les résultats associés aux données observées ainsi qu'une analyse de sensibilité, nous décrivons une étude par simulation en vue d'évaluer la différence entre les modèles de non-réponse ignorable et non ignorable.

4.1 Analyse des données

Consulter le tableau 2 pour une comparaison du modèle de non-réponse ignorable et du modèle de non-réponse non ignorable. Nous avons également inclus l'erreur-type numérique (ETN) qui est une mesure du degré de reproductibilité des résultats numériques; nous l'avons calculée par la méthode des moyennes par lot. Donc, cela ne nous gênerait pas que l'ETN soit petite comparativement aux estimations de Monte Carlo ou aux moyennes a posteriori. Pour les deux modèles, les ETN sont faibles, avec des valeurs relativement grandes pour le modèle de non-réponse non ignorable (proche de zéro dans les deux cas de toute façon), ce qui indique que les calculs sont reproductibles. Les moyennes a posteriori (MP) sont fort semblables pour les deux modèles. Les écarts-types a posteriori (ETP) sont plus grands pour le modèle de non-réponse non ignorable que

Alors, en posant que $d = 3!n!(rc - 1)!$ et $e = 3!n!(r - 1)!(c - 1)!$, la vraisemblance marginale pour le modèle de non-réponse ignorable (IG) est

$$p_{IG}(Y_1) = \left\{ \begin{array}{l} \text{association} \\ d \sum_{Y_{(1)} \in C} \prod_{s,j,k} (\pi_s p^{j,k})^{y_{sjk}} / y_{sjk}! \} dp dp, \\ \text{association} \\ e \sum_{Y_{(1)} \in C} \prod_{s,j,k} \{ (\pi_s q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}! \} p \pi d q_{1j} d q_{2k}, \\ \text{pas d'association,} \\ p_{NIG}(Y_1) = \end{array} \right.$$

et en posant que $\Omega_a = (p, \pi, \mu, \tau)$ et $\Omega_{na} = (q_1, q_2, \pi, \mu, \tau)$, la vraisemblance marginale pour le modèle de non-réponse non ignorable (NIG) est

$$p_{NIG}(Y_1) = \left\{ \begin{array}{l} \text{association} \\ d \sum_{Y_{(1)} \in C_a} \prod_{s,j,k} (\pi_{sjk} p^{j,k})^{y_{sjk}} / y_{sjk}! \} g(\pi, \mu, \tau) d\Omega_a, \\ \text{pas d'association,} \\ e \sum_{Y_{(1)} \in C_{na}} \prod_{s,j,k} \{ (\pi_{sjk} q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}! \} g(\pi, \mu, \tau) d\Omega_{na}, \\ \text{pas d'association,} \end{array} \right.$$

Dans l'ensemble C , la sommation demande beaucoup de calculs, parce qu'il existe de nombreux points $Y^{(1)} \in C$ (autrement dit, nous devons calculer la somme sur la totalité de ces points). Pour éviter ce problème, nous commençons par calculer la somme sur C analytiquement, puis nous obtenons le reste par intégration par la méthode de Monte Carlo.

Pour le modèle ignorable, il est facile de démontrer que

$$p_{IG}(Y_1) = \left\{ \begin{array}{l} \text{association} \\ a = \frac{n+1}{3!n!} (rc - 1)! \\ b = \frac{n+1}{3!n!} (n+r-1)!(c-1)! \prod_{j=1}^r y_{1j}! \prod_{f=1}^f \prod_{k=1}^k y_{1jk}! \end{array} \right.$$

où n est le nombre total d'individus dans le tableau entier. Nous décrivons l'estimation de $p_{NIG}(Y_1)$ à l'annexe B.

Cependant, nous notons qu'un test d'ignorabilité ou de non-ignorabilité est subtil, parce que nous supposons que nous ne disposons d'aucune information au sujet de l'ignorabilité ou de la non-ignorabilité. Par contre, notre modèle de non-réponse non ignorable est une généralisation de notre modèle de non-réponse ignorable. Nous pensons donc que le test d'association sous le modèle de non-réponse ignorable ou sous le modèle de non-réponse non ignorable est fiable.

Enfin, nous soulignons que le facteur de Bayes peut être sensible aux spécifications a priori, particulièrement parce qu'il n'existe pas suffisamment de données pour estimer les paramètres soumis au test; voir Sinharay et Stern (2002) pour une discussion intéressante des modèles emboîtés. Nous avons étudié la sensibilité du facteur de Bayes à la spécification de α_0 et de β_0 dans (17); consulter la section 3.5 et le tableau 6. Cet exercice est utile, car il s'agit d'un prior important dans notre modèle de non-réponse non ignorable. Cependant, la comparaison principale est un test d'absence d'association réalisé séparément sous le modèle de non-réponse ignorable et sous le modèle de non-réponse non ignorable. Le paramètre τ entre uniquement dans le modèle de non-réponse non ignorable et possède le même prior sous l'hypothèse d'association et d'absence d'association.

3.5 Spécification de α_0 et β_0

La spécification des hyperparamètres α_0 et β_0 dans $\tau \sim \text{Gamma}(\alpha_0, \beta_0)$ est un point essentiel de notre méthode; voir (12). Elle est importante, parce que nous utilisons cette technique pour rendre le modèle de non-réponse ignorable robuste; nous faisons une analyse de sensibilité plus loin. Notons que $E(\tau) = \alpha_0/\beta_0$; donc, si $\alpha_0 > \beta_0$, le modèle de non-réponse non ignorable sera semblable au modèle de non-réponse ignorable. Supposons que nous puissions observer un échantillon aléatoire $\tau^{(1)}, \dots, \tau^{(M)}$ tiré de la loi $\text{Gamma}(\alpha_0, \beta_0)$. Alors, nous pouvons utiliser une méthode simple (par exemple, la méthode des moments) pour estimer α_0 et β_0 . Comment pouvons-nous obtenir un échantillon correspondant à $\text{Gamma}(\alpha_0, \beta_0)$? L'échantillonneur de Gibbs donné en (8) pour le modèle de non-réponse ignorable produit les valeurs imputées pour les effectifs de cellule manquants. Nous avons imputé les effectifs de cellule manquants M fois, $M = 1000$; soit $n_{(h)}^{ijk} \equiv y_{1jk}^{ijk}$ et $u_{(h)}^{ijk}$, $s = 2, 3, 4$, $h = 1, \dots, M$, les effectifs de cellule manquants. Alors, pour chaque valeur de h , nous ajustons le modèle de non-réponse non ignorable sans la spécification a priori (12),

$$(n_{111}^{(h)}, \dots, n_{1rc}^{(h)}, n_{411}^{(h)}, \dots, n_{4rc}^{(h)}) | \pi, p \sim \text{Multinomial}\{n, (\pi_{111}, \dots, \pi_{4rc})\}.$$

Souignons que, dans (10), les paramètres p_k et π_{sjk} ne sont pas identifiables. Manifestement, pour estimer p_k , nous avons besoin de connaître y_{jk} , mais nous ne connaissons que y_{jk} . En outre, pour estimer π_{sjk} , nous devons connaître y_{sjk} , $s = 2, 3, 4$. Donc, les y_{sjk} , $s = 2, 3, 4$ ne sont pas identifiables non plus. Appliquer aux y_{sjk} des priors appropriés usés informatifs idéal, mais il ne s'agit pas d'une solution pratique. Si l'on utilise un modèle de non-réponse ignorable (c'est-à-dire $\pi_{sjk} = \pi_s$), alors tous les paramètres peuvent être identifiés. Par conséquent, une solution raisonnable consiste à essayer de lier les π_{jk} sur (j, k) en se servant d'une caractéristique commune. Si les π_{jk} proviennent d'une distribution commune dont les paramètres sont « connus », nous pourrions les estimer. Autrement dit, nous devons essayer de « renforcer l'information par emprunt », comme dans l'estimation sur petits domaines. Cela nous permettra d'estimer $y^{(1)}$ qui, à son tour, facilitera l'estimation des p_{jk} et π_{sjk} .

Pour les π_{sjk} , nous « centrons » le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable. Plus précisément, nous supposons que $\pi_{jk} | \mu, \tau \sim \text{Dirichlet}(\mu_1, \mu_2, \mu_3, \mu_4)$, où α_0 et β_0 doivent être spécifiés; sans aucune information au sujet de α_0 et β_0 , nous devons utiliser de nouveau les

indépendants avec $p(\mathbf{u}) = 1, \mu_s \geq 0, s = 1, 2, 3, 4$, $\sum_{s=1}^4 \mu_s = 1, \tau \sim \text{Gamma}(\alpha_0, \beta_0), \tau \geq 0$, (12)

naturel de procéder consiste à essayer d'utiliser certaines données déjà observées.

Plus précisément, nous prenons a priori que μ et τ sont

choisir une densité a priori pour τ telle que le modèle de information au sujet de la non-ignorabilité, il est naturel de choisir ce dernier prudemment. En l'absence de toute l'inférence peut être sensible au choix de τ , et il convient si τ est grand, les π_{jk} seront fort semblables. Donc, Par exemple, si τ est petit, les π_{jk} seront fort différentes et ignorable par rapport au modèle de non-réponse non renseigné sur la proximité du modèle de non-réponse non ignorable (11), le paramètre τ nous

$$\pi_{sjk} \geq 0, \sum_{s=1}^4 \pi_{sjk} = 1, \quad (11)$$

$y^{(1)}$ par

$$\pi(p, \pi, \mu, \tau, y^{(1)} | y_1) \propto \left\{ \prod_{s=1}^4 \prod_{r,c} \pi_{sjk}^{y_{sjk}} \right\} \frac{y_{sjk}!}{\pi_{sjk}^{y_{sjk}}} \quad (14)$$

À l'annexe A, nous montrons comment ajuster le modèle de non-réponse non ignorable pour obtenir une inférence appropriée en utilisant l'échantillonnage de Gibbs.

3.4 Facteur de Bayes : Tests d'association et de non-ignorabilité

Nous élaborons un test en vue de vérifier l'association entre DMO et RF. Ce test est une évaluation de l'hypothèse selon laquelle $p_{jk} = q_{1j}q_{2k}$, $j = 1, \dots, r$, $k = 1, \dots, c$, et $\sum_{j=1}^r q_{1j} = 1$ et $\sum_{k=1}^c q_{2k} = 1$. Nous utilisons le facteur de Bayes, c'est-à-dire le ratio des vraisemblances marginales sous deux scénarios (à savoir association contre absence d'association). Souignons que nous observons y_1 , mais que $y^{(1)}$ est un ensemble de variables latentes, de sorte que chaque probabilité marginale est simplement la probabilité que y_1 soit la valeur observée de X_1 , ce que nous notons par $p(y_1)$.

Nous fixons

$$C = \left\{ \begin{aligned} & y^{(1)} : \sum_{k=1}^c y_{2jk} = n_j, j = 1, \dots, r; \\ & \sum_{j=1}^r y_{3jk} = v_k, k = 1, \dots, c; \sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w \end{aligned} \right\}.$$

Statistique Canada, No 12-001-XPB au catalogue

Autrement dit, il n'existe aucune dépendance à l'égard de la

situation de cellule d'un individu.

Alors, la fonction de vraisemblance augmentée pour

$$p, \pi, y^{(1)} | y_1, n_0, n, v, w \text{ est}$$

$$g(p, \pi, y^{(1)} | y_1, n_0, n, v, w) \propto \left[\prod_{s=1}^4 \pi_s^{y_s} \right] \left[\prod_{s=1}^4 \prod_{r=1}^4 \prod_{k=1}^4 \prod_{j=1}^4 \frac{p_{y_{sjk}}}{y_{sjk}!} \right], \quad (4)$$

sous les contraintes $\sum_{j=1}^4 \sum_{k=1}^4 y_{jk} = n_0$, $\sum_{k=1}^4 y_{jk} = n_j$, $\sum_{j=1}^4 y_{jk} = v_k$, $k = 1, \dots, c$, et $\sum_{j=1}^4 \sum_{k=1}^4 y_{jk} = w$.

L'équation (4) possède trois caractéristiques intéressantes. Au premier lieu, sous l'hypothèse d'ignorabilité, la fonction de vraisemblance est subdivisée en deux éléments, l'un contenant les π_s uniquement et l'autre, les

p_{jk} , et les inférences au sujet de ces deux paramètres sont indépendantes. En deuxième lieu, l'inférence au sujet de π_s est fondée uniquement sur les $y_{s\cdots}$ observés (c'est-à-dire que

les statistiques suffisantes pour π_1, π_2, π_3 et π_4 sont essentiellement les proportions de cas dans les premier, deuxième, troisième et quatrième tableaux, respectivement).

En troisième lieu, sous le modèle de non-réponse ignorable, les n_j et les v_k contiennent de l'information au sujet de

p_{jk} . Il ne contient aucune information au sujet des p_{jk} . est facile de le démontrer; en notant l'ensemble $\{(y_2, y_3,$

$y_4) : \sum_{k=1}^4 y_{2k} = n_2, \sum_{j=1}^4 y_{3k} = v_k, k = 1, \dots, c$, d'après (4)

$$\sum_{s=1}^4 \prod_{r=1}^4 \prod_{k=1}^4 \prod_{j=1}^4 \frac{p_{y_{sjk}}}{y_{sjk}!} = w! \prod_{j=1}^4 \frac{n_j!}{\left\{ \sum_{k=1}^4 p_{jk} \right\}!} \prod_{r=1}^4 \prod_{k=1}^4 \prod_{j=1}^4 \frac{v_k!}{\left\{ \sum_{j=1}^4 p_{jk} \right\}!} \prod_{r=1}^4 \prod_{k=1}^4 \prod_{j=1}^4 \frac{w!}{\left\{ \sum_{j=1}^4 p_{jk} \right\}!}.$$

Enfin, pour les paramètres π , nous prenons

$$\pi \sim \text{Dirichlet}(1, \dots, 1), \pi_s \geq 0, \sum_{s=1}^4 \pi_s = 1, \quad (5)$$

Notons qu'il s'agit d'une densité de probabilité uniforme dans un espace quadridimensionnel et qu'il n'y a pas d'hyperparamètres dans ce modèle. Donc, pour le modèle de non-réponse ignorable, si nous combinons (2) et (5), la densité a priori conjointe est

$$g_1(p, \pi) \propto 1, p_{jk} \geq 0, \sum_{j=1}^4 \sum_{k=1}^4 p_{jk} \geq 0, \sum_{s=1}^4 \pi_s = 1, \quad (6)$$

ce qui est pertinent.

Enfin, en combinant la fonction de vraisemblance (4) et la densité a priori conjointe (6) par la voie du théorème de

3.3 Modèle de non-réponse non ignorable

De toute évidence, les paramètres p et π sont identifiables et estimables. En outre, notons que, dans (8), y_4 est une variable latente et qu'elle ne contribue pas à l'inférence au sujet de p . Elle facilite plutôt le calcul en fournissant un échantillonneur de Gibbs simple. Cependant, nous soulignerons que l'information contenue dans y_4 par la voie de w , est importante sous un modèle de non-réponse non ignorable.

$$y_4 | p, w, y^{(4)} \sim \text{Multinomial}(w, p), \quad (8)$$

$$y_{3k} | p, v_k, y^{(3)} \sim \text{Multinomial}(v_k, q_{(2)}^k), k = 1, \dots, c,$$

$$y_{2j} | p, n_j, y^{(2)} \sim \text{Multinomial}(n_j, q_{(1)}^j), j = 1, \dots, r,$$

$$p | y \sim \text{Dirichlet}(y_{11} + 1, \dots, y_{rc} + 1),$$

A posteriori, p et π sont indépendants. L'inférence au sujet de π est facile, car $\pi | y_1, y^{(1)} \sim \text{Dirichlet}(y_{1\cdot} + 1, \dots, y_{4\cdot} + 1)$, qui est indépendant de $y^{(1)}$. L'inférence au sujet de p peut s'obtenir en utilisant un simple échantillonneur de Gibbs, car, si nous posons que $q_{(1)}^j = p_{jk} / \sum_{k=1}^4 p_{jk}$ et $q_{(2)}^k = p_{jk} / \sum_{j=1}^4 p_{jk}$, les probabilités conditionnelles sont

$$\pi(p, \pi, y^{(1)} | y_1) \propto \left[\prod_{s=1}^4 \pi_s^{y_s} \right] \left[\prod_{s=1}^4 \prod_{r=1}^4 \prod_{k=1}^4 \prod_{j=1}^4 \frac{p_{y_{sjk}}}{y_{sjk}!} \right]. \quad (7)$$

Bayes, nous obtenons la densité a posteriori conjointe des paramètres π, p et $y^{(1)}$

$$g(p, \pi, y^{(1)} | y_1, n_0, n, v, w) \propto \left[\prod_{s,j,k} \pi_{y_{sjk}} \frac{p_{y_{sjk}}}{y_{sjk}!} \right] \left[\prod_{r,c} \prod_{k=1}^4 p_{y_{rk}}^{y_{rk}} \right], \quad (10)$$

pour $p, \pi, y^{(1)} | y_1$ est semblance. Ici, la fonction de vraisemblance augmentée

Ensuite, nous avons besoin de la fonction de vraisemblance augmentée de π_{jk} ; la méthodologie est la même.

Han et Choi (2002). Nous pouvons aussi prendre π_j ou π_k ignorables. Il s'agit d'une extension du modèle de Nandram, selon laquelle les données manquantes ne sont pas caractéristiques (c'est-à-dire classifications ligne et colonne) appartenant à l'un des quatre tableaux dépend des deux l'hypothèse (9) précise que les probabilités qu'un individu

$$\sim \text{Multinomial}(1, \pi^{jk}). \quad (9)$$

$$I^{jk} = 1, I^{j'k'} = 0, j \neq j', k \neq k', \pi^{jk}$$

preons

Pour les données manquantes non ignorables, nous

celui dont les composantes correspondent aux quatre tableaux.

Soit p_{jk} la probabilité qu'un individu appartienne à la cellule (j, k) du tableau $r \times c$, et soit π_{sjk}^r la probabilité qu'un individu appartienne au s^e tableau, sachant la situation de cette cellule (j, k) . Pour le modèle de non-réponse ignorable, π_{sjk}^r dépend au moins d'une valeur de j ainsi que de k . Nous supposons aussi que p est le vecteur dont les composantes sont $\{\pi_{sjk}^r; s=1, \dots, 4\}$, $j=1, \dots, r$, $k=1, \dots, c$.

Alors, nous prenons

$$I \mid p \sim \text{Multinomiale}\{1, p\}, \quad (1)$$

où $\sum_{j=1}^r \sum_{k=1}^c p_{jk} \geq 0$, $j=1, \dots, r$, $k=1, \dots, c$. Pour les paramètres p , nous prenons

$$p \sim \text{Dirichlet}(1, \dots, 1), \quad p_{jk} \geq 0, \quad \sum_{j=1}^r \sum_{k=1}^c p_{jk} = 1. \quad (2)$$

Désormais, nous utiliserons la notation qu'un vecteur de dimension k , $x \sim \text{Dirichlet}(c)$ signifie $f(x) = \prod_{j=1}^k x_j^{c_j-1} / \prod_{j=1}^k \Gamma(c_j)$, où $D_k(c) = \{x_j \geq 0, \sum_{j=1}^k x_j = 1\}$, et la fonction de Dirichlet avec $c_j > 0$, $\sum_{j=1}^k c_j = 1$.

Les hypothèses (1) et (2) sont les mêmes pour les modèles de non-réponse ignorable et non ignorable, et sont les hypothèses types lorsqu'il n'y a pas de données manquantes.

Soit $y_{sjk} = \sum_{l=1}^n I_{jkl}^s J^{sl}$, $s=1, 2, 3, 4$ les effectifs de cellule pour les quatre cas. Ici, les valeurs y_{1jk} sont observées et les valeurs y_{sjk} , $s=2, 3, 4$ manquent (c'est-à-dire, variables latentes). Pour y_{1jk} nous savons que $\sum_{k=1}^c y_{1jk} = n_0$, le nombre d'individus pour lesquels les données sont complètes; pour y_{2jk} , nous savons que $\sum_{k=1}^c y_{2jk} = n_1$, où les totaux de ligne n_j , $j=1, \dots, r$ sont observés; pour y_{3jk} , nous savons que $\sum_{j=1}^r y_{3jk} = v_k$, où les totaux de colonnes v_k , $k=1, \dots, c$ sont observés, et pour y_{4jk} nous savons que $\sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w$. Partout, nous supposons que toute inférence est faite sachant n_0, n_1, v et w , et nous supprimerons cette notation chaque fois que cela est sous-entendu. Lorsque cela est commode, nous utiliserons les notations telles que $\sum_{s,j,k} \pi_{sjk}^r$ et $\sum_{s,j,k} \pi_{sjk}^r \equiv \prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}^r$, $Y^{(1)} = (y_1, y_2, y_3, y_4)$, $Y^{(2)} = (y_1, \dots, y_r, k=1, \dots, c)$, $s=1, 2, 3, 4$. En outre, $\sum_{s,j,k} y_{sjk} = n$. Nous utiliserons aussi $y_{s \cdot} = \sum_{j,k} y_{sjk}$, $y_{j \cdot} = \sum_{k} y_{sjk}$ et $y = (y_1, y_2, y_3, y_4)$.

3.2 Modèle de non-réponse ignorable

Pour le modèle de non-réponse ignorable, nous prenons

$$J^r \mid \pi \sim \text{Multinomiale}\{1, \pi\}. \quad (3)$$

où $D_r(\cdot)$ représente la fonction de Dirichlet avec r composantes, etc.; voir la section 3.1 pour la notation. Puis, nous choisissons chacune des composantes de α, β et γ comme étant κ (par exemple, dans $p_{\text{as}}(u)$ et $p_{\text{as}}(n)$, $\kappa=1$). Nous pouvons étudier la sensibilité du choix des lois a priori qui fonction de κ . Ici $\kappa=1$ correspond aux lois a priori qui sont habituellement utilisées dans le modèle Dirichlet-Multinomiale et $\kappa=0,50$ correspond à la loi a priori de Jeffreys. Donc, nous avons choisi $\kappa=0,25, 0,5, 1,0, 1,5$ et $2,3$ et les facteurs de Bayes correspondants (échelle logarithmique) sont 4,7, 3,6, 3,4, 3,9, 4,7 et 6,6. Par conséquent, le facteur de Bayes est sensible au choix des lois a priori, mais pas excessivement. Naturellement, s'il existe des données a priori informatives, pour lesquelles la valeur de κ est fort grande, le problème est différent.

La statistique du chi-carré de Pearson est dominée par les cellules (3, 1) et (3, 3) pour lesquelles les carrés des résidus de Pearson sont 4,61 et 6,15, respectivement (la statistique du chi-carré observée est 12,7). Il est intéressant de noter que le facteur de Bayes a tendance à lisser cet effet. Nous avons regroupé les deux catégories, ostéopathe et ostéopathe-rose, en une seule. Pour ce tableau de contingence 2×3 , la valeur de la statistique du chi-carré est 1,7 pour deux degrés de liberté avec une valeur p de 0,42. Les vraisemblances marginales sont $p_{\text{as}}(u) = -28,2$ et $p_{\text{as}}(n) = -32,0$, ce qui donne un logarithme du facteur de Bayes de -3,81. Par conséquent, les deux tests donnent à penser qu'il n'existe pas d'association pour ce tableau 2×3 . Donc, si l'on s'en tient à ces données, il est difficile de croire qu'il existe une association entre la DMO et le RF. La question qui se pose maintenant est celle de savoir si cette conclusion change quand on tient compte des données incomplètes.

3. Méthodologie et modèles de non-réponse

Premièrement, nous décrivons la notation. Deuxièmement, nous décrivons le modèle de non-réponse ignorable. Troisièmement, nous construisons un modèle de non-réponse non ignorable en étendant le modèle de non-réponse ignorable. Quatrièmement, nous discutons du facteur de Bayes. Enfin, nous décrivons la façon de spécifier une loi a priori importante.

3.1 Notation

Pour un tableau de contingence $r \times c$, soit $I^{jk} = 1$ si j^e ligne et dans la k^e colonne, et 0 autrement. En outre, soit $J^{sl} = 1$ si le l^e individu se trouve dans le tableau s ($s=1$: cas complets; $s=2$: tableau avec les totaux de ligne; $s=3$: tableau avec les totaux de colonne; $s=4$: tableau avec les individus non classés), et $J^{sl} = 0$ autrement, $s=1, 2, 3, 4$ avec $\sum_{s=1}^4 J^{sl} = 1$. Le vecteur $J^r = (J^{1r}, J^{2r}, J^{3r}, J^{4r})^r$ est

discussion. En outre, le taux de réponse aux questions sur le revenu est habituellement faible.

Nous avons examiné de plus près les données pour les cas complets. Nous avons ajusté un modèle Dirichlet-Multinomial avec association et un autre sans association. Le modèle avec association est $\mathbf{u} | \mathbf{p} \sim \text{Multinomial}(\mathbf{n}, \mathbf{p})$ et $\mathbf{p} \sim \text{Dirichlet}(1, \dots, 1)$. Notons que, par absence d'association, nous entendons que $p_{jk} = p_{(j)}^{(k)} p_{(k)}^{(j)}$, $j = 1, \dots, r$, $k = 1, \dots, c$, où $\sum_{j=1}^r p_{(j)}^{(k)} = 1$ et $\sum_{k=1}^c p_{(k)}^{(j)} = 1$. Donc, pour le modèle sans association, $\mathbf{u} | \mathbf{p} \sim \text{Multinomial}(\mathbf{n}, \mathbf{p})$, $\mathbf{p}^{(1)} \sim \text{Dirichlet}(1, \dots, 1)$, et indépendamment $\mathbf{p}^{(2)} \sim \text{Dirichlet}(1, \dots, 1)$, où $\mathbf{p}^{(1)}$ et $\mathbf{p}^{(2)}$ possèdent r et c composantes, respectivement. Il est facile de montrer que la vraisemblance marginale avec association (as) est $p_{\text{as}}(\mathbf{u}) = (rc - 1)! n! / ((n + rc - 1)!)^2$ et que la vraisemblance marginale sans association (nas) est

$$p_{\text{nas}}(\mathbf{u}) = \frac{(r - 1)!(c - 1)!}{(n + r - 1)!(n + c - 1)!} \frac{(rc - 1)!}{(n + r - 1)!(n + c - 1)!} \prod_{j=1}^r n_{.j}! \prod_{k=1}^c n_{.k}!$$

Revenons à nos données du tableau I. Sous l'hypothèse d'indépendance (c'est-à-dire pas d'association), la statistique du chi-carré observée est 12,7 pour quatre degrés de liberté avec une valeur p de 0,013 et nous rejetons l'hypothèse d'absence d'association. Sur l'échelle logarithmique, les vraisemblances marginales sont $p_{\text{nas}}(\mathbf{u}) = -46,2$ et $p_{\text{as}}(\mathbf{u}) = -49,6$ menant à un logarithme du facteur de Bayes de 3,40 pour la preuve d'une absence d'association relativement à une association. Par conséquent, alors que le test du chi-carré indique fortement qu'il faut rejeter l'hypothèse d'absence d'association, le logarithme du facteur de Bayes indique fortement qu'il faut l'accepter. Donc, les données concernant l'absence d'association sont contradictoires. Voir Mirkin (2001) pour une revue des interprétations de la statistique du chi-carré en tant que mesure d'association ou d'indépendance.

À quel point le facteur de Bayes est-il sensible au choix des lois a priori? En premier lieu, notons que la densité a priori que toute personne raisonnable choisirait pour résoudre le présent problème est la loi de Dirichlet. Pour le modèle avec association, nous choisissons comme lois a priori $\mathbf{p} \sim \text{Dirichlet}(\gamma)$, et pour le modèle sans association, $\mathbf{p}^{(1)} \sim \text{Dirichlet}(\mathbf{a})$ et indépendamment $\mathbf{p}^{(2)} \sim \text{Dirichlet}(\mathbf{b})$. Soit $n_{.j}^{(1)} = \sum_{k=1}^c n_{jk}^{(1)}$, $j = 1, \dots, r$ et $n_{.k}^{(2)} = \sum_{j=1}^r n_{jk}^{(2)}$, $k = 1, \dots, c$. Alors, il est facile de montrer que le facteur de Bayes (FB) pour un test d'association contre absence d'association est

$$\text{FB} = \frac{D^{rc}(\mathbf{u} + \gamma) / D^r(\mathbf{u}) D^c(\mathbf{u})}{D^{rc}(\mathbf{u}) / D^r(\mathbf{u}) D^c(\mathbf{u})} \frac{D^{rc}(\gamma) / D^r(\mathbf{a}) D^c(\mathbf{b})}{D^{rc}(\mathbf{u}) / D^r(\mathbf{a}) D^c(\mathbf{b})},$$

au tableau I sous forme de tableau de contingence 3×3 de proportion d'individus pour chaque niveau, ou cellule, DMO-RF et de vérifier l'hypothèse d'une association entre DMO et RF. Dans la NHANES III, le taux de réponse augmente jusqu'à l'âge de 20 ans, puis se stabilise après cet âge; la race, le sexe et les poids de sondage jouent un rôle mineur (voir Nandram et Choi 2005). Donc, ici, nous supposons que les seules données disponibles sont les quatre tableaux regroupant la DMO et le RF, et nous élaborons une méthode applicable à cette situation.

Tableau 1
Classification de la densité minérale osseuse (DMO) et du revenu familial (RF) pour 2 998 femmes blanches ayant au moins 20 ans (20+)

	RF		DMO	
	0	1	0	1
Manquant	2	1	290	284
Somme	1 330	135	69	577
0	621	290	131	117
1	260	131	30	18
2	93	30	156	266
Manquante	430	607	276	45
Somme	1 430	685	2 998	2 998

Nota: DMO: 0 (< 0,82g/cm²; normale), 1 (> 0,64, ≤ 0,82g/cm²; ostéopénie), 2 (≤ 0,64g/cm²; ostéoporose); RF: 0 (< 20 000 \$), 1 (≥ 20 000 \$, < 45 000 \$), 2 (≥ 45 000 \$); la DMO est mesurée uniquement pour les femmes de 20 ans et plus.

Il est difficile d'évaluer une association entre DMO et RF quand la classification est incomplète (c'est-à-dire données manquantes) pour de nombreux individus. Comme il est discuté dans la littérature ne portant pas nécessairement sur la NHANES III, il existe plusieurs variables éventuellement conventionnelles importantes, comme l'âge, l'usage du tabac, l'apport alimentaire de calcium, l'ostrogène/thérapie substitutive, l'activité physique, le niveau de scolarité, l'état de santé et la consommation d'alcool (voir Canry, Baudoin et Fardellone 2000). Selon Farahmand, Persson, Michaëlsson, Baron, Parker et Ljunghall (2000), chez les femmes ménopausées de 50 à 81 ans de six comités de la Suède, un revenu du ménage élevé est associé à une diminution du risque de fracture de la hanche. Au moyen de l'ensemble complet de données provenant de la NHANES III, Lauderdale et Rathouz (2003) étudient la régression de la teneur minérale osseuse sur les indicateurs économiques (par exemple, le niveau de scolarité et le ratio pauvre/revenu). Ils font une correction pour tenir compte d'autres facteurs comme l'âge, la taille et le poids. Ils concluent que la densité osseuse ne reflète pas les conditions économiques aussi fortement que la teneur minérale osseuse. Malheureusement, les auteurs de tous ces travaux n'abordent pas la question de la non-ignorabilité des données manquantes; ces dernières ne font l'objet d'aucune

1.2 Méthodes connexes

Notre méthode diffère de celle de Rubin, Stern et Vehovar (1995). Nous partons de l'approche de Nandram et Choi (2002 a, b) selon laquelle un paramètre γ centre (peut-être considéré comme un indice) le modèle de non-réponse non ignorable sur le modèle de non-réponse. Si $\gamma = 1$, le modèle de non-réponse non ignorable coïncide avec le modèle de non-réponse ignorable et, donc, le modèle de non-réponse non ignorable « dégenère » en le modèle de non-réponse ignorable quand $\gamma = 1$; voir aussi Forster et Smith (1998). Cette approche est utile, car le modèle de non-réponse ignorable contient le modèle de non-réponse ignorable en tant que cas particulier, exprimant de ce fait l'incertitude quant à l'ignorabilité. Draper (1995) a donné à cette approche le nom d'*extension continue de modèle* et a recommandé son utilisation de préférence à une extension discrète de modèle (c'est-à-dire mélanges finis) dans la mesure du possible. Nous appelons simplement l'extension continue de modèle un modèle à *facteur d'extension*. Nandram et Choi (2002 a, b) obtiennent le centrage en prenant $\gamma | v \sim \text{Gamma}(v, v)$ avec $E(\gamma | v) = 1$, $\text{var}(\gamma | v) = 1/v$.

Nandram et Choi (2002 a) analysent des données binaires sur les crimes domestiques provenant de la National Crime Survey et, dans Nandram et Choi (2002 b), des données binaires sur les visites chez le médecin provenant de la National Health Interview Survey. Alors que Nandram et Choi (2002 a) contient plus de comparaisons, Nandram et Choi (2002 b) comprend un plus grand nombre d'analyses de sensibilité. Nandram, Han et Choi (2002) décrivent deux modèles bayésiens hiérarchiques, soit un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable, pour l'analyse des données de dénombrement provenant de plusieurs régions, les dénombrements étant décrits pour chaque région par une loi multinomiale. Dans tous ces travaux, la question de l'association est sens objet, car il n'existe qu'une seule variable nominale. L'approche de Nandram et Choi (2002 a, b) est séduisante, mais elle ne peut s'appliquer directement au problème considéré ici du tableau de contingence $r \times c$. Plus précisément, Nandram et Choi (2002 a, b) n'ont eu besoin que d'un seul paramètre de centrage. Pour étendre leur méthode, nous avons besoin de rc paramètres de centrage. La distribution de chacun de ces paramètres doit être centrée sur la valeur 1 pour permettre la dégénération en le modèle de non-réponse ignorable. Des contraintes d'ingélaté doivent également être incluses dans le modèle de non-réponse non ignorable. Par conséquent, bien que l'idée soit intéressante, la méthodologie nécessaire pour appliquer les travaux de Nandram et Choi (2002 a, b) dépasse de loin le cadre du présent article.

2. Données sur la densité minérale osseuse et le revenu familial

Nandram, Liu, Choi et Cox (2005) étendent les travaux de Nandram, Han et Choi (2002) dans deux directions importantes afin a) de considérer plusieurs tableaux de contingence à deux variables au lieu de tableaux à une seule variable et b) d'élaborer une méthode permettant d'étudier l'association entre les deux variables catégorielles. Nandram, Liu, Choi et Cox (2005) analysent les données sur la relation entre la densité minérale osseuse (DMO) et l'âge recueillies pour 35 comtés dans le cadre de la troisième National Health and Nutrition Examination Survey. Dans chaque comté, les données sont ventilées en deux catégories d'âge et trois catégories de DMO (autrement dit, il existe 35 tableaux de contingence 2×3). Il convient de souligner que l'âge est observé pour chaque individu, mais que les valeurs de la DMO manquent pour un grand nombre d'entre eux. Donc, pour chaque comté, il existe un tableau contenant les cas complets et un tableau contenant les totaux de ligne (c'est-à-dire les cas pour lesquels l'âge est connu, mais non la valeur de la DMO). Ici, l'objectif est d'étendre les travaux de Nandram, Liu, Choi et Cox (2005) à un tableau de contingence $r \times c$ général. Il s'agit d'un progrès important, puisque nous avons trois tableaux supplémentaires (un tableau avec la classification ligne unique-mentaire) avec la classification colonne uniquement et un troisième ne contenant ni classification ligne ni classification colonne) au lieu d'un seul avec les totaux de ligne comme dans Nandram, Liu, Choi et Cox (2005).

Nous décrivons brièvement le tableau de contingence 3×3 de la densité minérale osseuse (DMO) et du revenu familial (RF). RF est une variable discrète comportant trois niveaux : faible, moyen et élevé. Bien que DMO soit une variable continue, l'Organisation mondiale de la santé l'a classée en trois niveaux : normal, ostéopénie et ostéoporose; voir Looker, Orwoll, Johnston, Lindsay, Wahner, Dunn, Calvo et Harris (1997, 1998). DMO est utilisée pour diagnostiquer l'ostéoporose, maladie qui se manifeste chez les femmes âgées et, dans la NHANES III, elle est évaluée chez des individus ayant au moins 20 ans (autrement dit, nous utilisons les données sur les femmes blanches uniquement, ayant plus de 20 ans et présentant des problèmes de santé chroniques). Parmi celles ayant participé à la phase d'examen physique, des données sur le RF ainsi que sur la DMO ont été recueillies auprès d'environ 62 %, des données sur la DMO uniquement auprès de 8 %, des données sur le revenu uniquement auprès de 29 %; enfin, aucune donnée sur le revenu ni sur la DMO n'ont été recueillies auprès de 1 %.

Analyse bayésienne des données catégoriques manquantes non ignorables : Une application à la densité minérale osseuse et au revenu familial

Balgobin Nandram, Lawrence H. Cox et Jai Won Choi¹

Résumé

Le problème que nous considérons nécessite l'analyse de données catégoriques provenant d'un seul tableau à double entrée avec classification partielle (c'est-à-dire avec non-réponses partielles et totales). Nous supposons qu'il s'agit de la seule information disponible. Une méthode bayésienne nous permet de modéliser divers scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ignorabilité. Nous construisons un modèle de non-réponse non ignorable que nous obtenons par extension du modèle de non-réponse ignorable au moyen d'une loi a priori dépendante des données; l'extension au modèle de non-réponse non ignorable rend le modèle de non-réponse ignorable plus robuste. Nous utilisons un modèle Dirichlet-Multinomial, corrigé pour la non-réponse, pour estimer les probabilités de cellule et un facteur de Bayes pour vérifier l'hypothèse d'association. Nous illustrons notre méthode à l'aide de données sur la densité minérale osseuse et sur le revenu familial. Une analyse de sensibilité nous permet d'évaluer l'effet du choix de la loi a priori dépendante des données. Nous comparons les modèles de non-réponse ignorable et non ignorable au moyen d'une étude par simulation et constatons qu'il existe des différences subtiles entre ces modèles.

Mots clés : Facteur de Bayes; statistique du chi-carré; fonction d'importance; simulation de Monte Carlo à chaînes de Markov; modèle Dirichlet-Multinomial; robuste; tableau de contingence à double entrée.

1. Introduction

En pratique, il est courant d'utiliser des tableaux de contingence à double entrée pour présenter les données d'enquête. Souvent, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Donc, les tableaux à double entrée contiennent à la fois des non-réponses partielles (cas où l'une des deux caractéristiques manque) et des non-réponses totales (cas où les deux caractéristiques manquent); voir Little et Rubin (2002, section 1.3) pour les définitions des trois mécanismes donnant lieu aux données manquantes (MCAR – manquent entièrement au hasard, MAR – manquent au hasard, MNAR – ne manquent pas au hasard). Il peut donc exister quatre types de tableau (un tableau où tous les cas sont complets et, éventuellement, trois tableaux contenant, respectivement, les cas avec classification (données) ligne uniquement, les cas avec classification colonne uniquement et les cas sans classification ligne ni colonne). Il se peut que l'on ne connaisse pas le mécanisme produisant les données manquantes. Donc, nous utilisons un modèle dans lequel la fonction de vraisemblance tient compte des différences entre les données observées et les données manquantes (c'est-à-dire données manquantes non ignorables); voir Rubin (1976), ainsi que Little et Rubin (2002) pour la relation entre l'ignorabilité ou la non-ignorabilité et les trois mécanismes de production des données manquantes. Comme la méthode bayésienne offre des avantages bien connus par rapport à

Nous supposons que les seules données dont dispose l'analyse sont les cas complets et les trois tableaux supplémentaires. Plus précisément, nous posons qu'il n'existe aucune donnée (provenant de covariables ou d'information a priori) sur la non-ignorabilité. Notre approche bayésienne ne tient pas compte des caractéristiques du plan de sondage (c'est-à-dire pas de poids de sondage, et pas de mise en grappe ni de stratification). Pour présenter les données d'enquête au public, il arrive qu'on supprime les données sur certaines caractéristiques par souci de commodité et pour assurer la protection des renseignements personnels. Nous sommes conscients que le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable pourraient l'un et l'autre être incorrects, s'ils ne tiennent pas compte de ces caractéristiques. Cependant, les paramètres du modèle de non-réponse ignorable peuvent être identifiés et estimés, et nous tirons parti de ce fait pour construire un modèle de non-réponse non ignorable qui est relié au modèle de non-réponse ignorable. En outre, dans le modèle de non-réponse ignorable, nous supposons que la non-réponse est produite selon un mécanisme MAR et que

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute road, Worcester MA 01609, Courriel : balmann@wpi.edu; Lawrence H. Cox et Jai Won Choi, Office of Research and Methodology, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782, Courriel : lgc9@cdc.gov, jwco1@cdc.gov.

Tableau A5
Nombres totaux selon la taille de la famille et du ménage pour l'échantillon completé imputé.
Fondé sur le modèle (4.9)

Taille de la famille		Taille du ménage	
1	2	3	4
187,9	85,7	26,3	12,7
23,0	309,2	48,6	5,7
23,2	43,0	56,7	7,7
3	21,9	48,7	327,2
4	2,0	5,2	24,1
≥ 5	240,0	301,1	426,3
Total			

Bibliographie

- Baker, S.G., et Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.
- Belby, L. (1995). Forbrukersundersøkelser. Vektmetoder, Weights methods, nonresponse correction and interviewer effect). Notat 95/18 Statistics Norway.
- Bjørnstad, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.
- Bjørnstad, J.F., et Skjold, F. (1992). Interval estimation in the presence of nonresponse. *The American Statistical Association 1992 Proceedings of the Section on Survey Research Methods*, 233-238.
- Bjørnstad, J.F., et Walsøe, H.K. (1991). Predictive likelihood in nonresponse problems. *The American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*, 152-156.
- Efron, B., et Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.
- Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (avec discussion). *Journal of the Royal Statistical Society B*, 60, 57-70.
- Greenlees, J.S., Reece, W.S. et Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- Hall, B.H., Cummings, C. et Schake, R. (1991). *TSP Reference Manual, Version 4.2A*, Palo Alto California: TSP International.
- Holt, D., et Smith, T.M.F. (1979). Post-stratification, *Journal of the Royal Statistical Society A*, 142, 33-46.
- Kellman, N., et Brunborg, H. (1995). *Household Projections for Norway, 1990-2020, Part I: Macrosimulation*, Rapport 95/12, Statistics Norway.
- Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- Little, R.J.A., et Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.
- McCullagh, P., et Nelder, J.A. (1991). *Generalized Linear Models*, 2^{ème} éd. London: Chapman & Hall.
- Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Statistics Norway (1990). *Survey of Consumer Expenditure 1986-88*. Official Statistics of Norway NOS B919.
- Statistics Norway (1996). *Survey of Consumer Expenditure 1992-1994*. Official Statistics of Norway NOS C317.
- Sæmstad, C.-E., Swensson, B. et Weteman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Tableau A1
Tailles de la famille et du ménage pour l'Enquête sur les dépenses de consommation de la Norvège de 1992, selon la région rurale ou urbaine. L'entrée supérieure est pour le groupe urbain

Taille de la famille		Taille du ménage					Taux de réponse	
		1	2	3	4	≥ 5	Total	Total
1	urbain	28	24	13	7	0	61	0,439
	rural	24	24	17	7	2	101	0,574
2	urbain	6	70	12	3	0	91	0,520
	rural	3	107	25	1	3	139	0,647
3	urbain	4	8	57	11	3	83	0,675
	rural	6	17	74	29	3	129	0,717
4	urbain	0	3	15	80	5	103	0,705
	rural	2	10	22	151	12	197	0,711
≥ 5	urbain	0	1	0	6	66	73	0,723
	rural	1	3	4	11	115	134	0,807
Total urbain		38	106	91	102	74	411	0,601
Total rural		67	161	138	199	135	700	0,690

donc

$$\lambda^x = \sum_{z=0}^z \frac{p(R=0|x,z)}{m^{xz}(z)} - (m^x + m^{xz}).$$

Il découle de (A1) que p_{yx} satisfait la relation suivante :

$$p_{yx} = \frac{\left(m^x + m^{xz} - \sum_{z=0}^z m^{xz} \frac{p(R=0|I=y, x, z)}{p(R=0|I=y, x, z)} \right)}{n_{xy}}. \quad (A2)$$

Les valeurs imputées sont données par, tiré de (4.2),

$$n_{xy}^* (z) = m^{xz} (z) p_{yx} \frac{p(R=0|I=y, x, z)}{p(R=0|I=x, z)}$$

et, d'après (A2),

$$p_{yx} = n_{xy} \left(m^x + m^{xz} - \sum_{z=0}^z \frac{p_{yx} n_{xy}^* (z)}{n_{xy}} \right) / \left(m^x + m^{xz} - \frac{p_{yx} n_{xy}^*}{n_{xy}} \right)$$

ou bien, de façon équivalente

$$p_{yx} (m^x + m^{xz}) - n_{xy}^* = n_{xy},$$

c'est-à-dire, $p_{yx} = \frac{n_{xy}^* + m^{xz}}{n_{xy} + m^{xz}}$. C.Q.F.D.

Nous déterminons λ^x par sommation sur y :

$$- \sum_{z=0}^z m^{xz} \frac{p(R=0|I=x, z)}{p(R=0|I=x, z)} - p_{yx} \lambda^x.$$

$$n_{xy} p_{yx} = \frac{\sum_{z=0}^z m^{xz} \frac{p(R=0|I=x, z)}{p(R=0|I=x, z)}}{m^{xz}(z)}$$

qui est équivalent à :

$$(A1) \quad \frac{p_{yx}}{n_{xy}} - \sum_{z=0}^z m^{xz} \frac{p(R=0|I=x, z)}{p(R=0|I=x, z)} + \lambda^x = 0$$

Nous utilisons la méthode de Lagrange et maximisons $G = \ell + \sum_{x=1}^5 \lambda^x (\sum_{y=1}^5 p_{yx} - 1)$.
Posons que les solutions sont p_{yx}^* (λ^x) et déterminons les λ^x tels que $\sum_y p_{yx}^* (\lambda^x) = 1$, pour tout x . Quelle que soit la façon dont les q_{yxz} sont paramétrisés, l'emv p_{yx}^* doit satisfaire, en résolvant les équations $\partial G / \partial p_{yx} = 0$,

$$\ell = \sum_x \sum_y n_{xy} p_{yx} + \sum_y \sum_x \sum_z n_{xy} p_{yxz} (z) q_{yxz} + \sum_x \sum_z m^{xz} \log P(R=0|x,z)$$

Comparativement aux probabilités de réponse estimées d'après le modèle RM2 (y, z) avec (3.1), nous voyons que le remplacement de la taille du ménage par la taille de la famille dans le groupe de non-répondants n'est pas une approximation satisfaisante. Donc, si l'on fait une comparaison 5.1, fondée sur le modèle saturé (4.9), ce dernier serait adopté officiellement surestime la probabilité de réponse représentatif, produirait une sous-estimation de H_2 . Les probabilités de réponse estimées seront le plus probablement biaisées si nous utilisons la taille de la famille à la place de la taille du ménage dans le groupe de non-répondants pour estimer les paramètres du modèle de réponse. Ce biais est un problème qui s'ajoute à celui mentionné antérieurement, à savoir que l'estimateur d'Horvitz-Thompson modifié produit des estimations semblables à celles produites par l'estimateur par expansion basées sur l'imputation et ne peut corriger pour les échantillons non représentatifs (problème qui se pose dans le cas de l'EDC depuis 1993). Toutefois, pour l'EDC de 1992, l'échantillon est asymétrique et contient une proportion trop élevée de familles de deux personnes, et l'ordre de grandeur de l'estimation de H_2 sera correcte par accident.

6. Conclusions

Nous avons étudié les problèmes de modélisation et de méthodologie que pose l'estimation du nombre total de ménages de diverses tailles en Norvège d'après les données de l'Enquête sur les dépenses de consommation (EDC) de la Norvège. Le problème principal est de savoir comment corriger le biais dû à la non-réponse non ignorable. La méthode d'estimation appliquée à l'heure actuelle pour l'EDC est l'utilisation d'un estimateur d'Horvitz-Thompson modifié qui comprend une correction pour la non-réponse par estimation des probabilités de réponse. Nous avons fondamentalement examiné deux approches fondées sur un modèle, à savoir un estimateur du maximum de vraisemblance et une poststratification basée sur l'imputation d'après la taille enregistrée de la famille. Avec un modèle de population correspondant à un modèle de groupe d'après la taille de la famille uniquement, ces deux estimateurs sont identiques. Ce modèle de groupe de famille, pour la taille du ménage et une fonction de lien logistique pour la probabilité de réponse avec la taille du ménage comme variable catégorique nominale semble donner de bons résultats pour notre problème d'estimation.

Lors de l'analyse des données de l'EDC de 1992, nous observons un biais de non-réponse important, particulièrement dans les estimations de H_1 et H , biais qui est

Annexe A1

Les données pour les régions rurales et urbaine sont présentées séparément au tableau A1.

Annexe A2

Théorème. Supposons que Y suit le modèle (3.1), c'est-à-dire que $P(Y = y | x, z) = p_{yx}$ est indépendant de z , mais que, par ailleurs, les p_{yx} sont complètement inconnues, la seule contrainte étant que $\sum_y p_{yx} = 1$, pour toutes les valeurs de x , pour tout k . Le mécanisme de réponse est paramétrisé arbitrairement, autrement dit aucune hypothèse n'est faite au sujet de $P(R = 1 | Y = y, x, z)$. Alors, les estimations du maximum de vraisemblance pour p_{yx} sont données par

$$p_{yx} = \frac{n_{xy}}{n_{xy}^* + m_{xy}^*}$$

Preuve. Soit $q_{yz} = P(R = 1 | Y = y, x, z)$. La vraisemblance est donnée par

Donc, nous pouvons considérer les poids comme étant une approximation de (3.5). Évidemment, (3.5) n'est possible que si l'on envisage un modèle de population, ce qui n'a pas été fait dans le cas de l'EDC. Le tableau 7 donne la distribution estimée des ménages d'après cet estimateur d'Horvitz-Thompson modifié pour l'EDC. L'Enquête sur la qualité du Recensement de 1990, appelée EEP 1990, compte 8 280 répondants et repose sur pratiquement la même définition du ménage que l'EDC. Le taux de réponse était de 95 %. Les estimations de H_y sont calculées par poststratification en fonction de la taille du ménage au recensement. Cependant, aucun effort n'a été fait pour corriger le biais de non-réponse éventuel en ce qui a trait à la taille réelle du ménage. L'EEP porte sur l'ensemble de la population. Le tableau 7 donne les estimations pour le groupe de personnes de 0 à 79 ans calculées selon la même méthode de poststratification que pour l'EEP.

Le tableau 7 présente aussi les estimations basées sur l'étude de projection des ménages réalisée par Keilman et Brunborg (1995). Cette étude simule la structure des ménages pour la période allant de 1990 à 2020. Les sources sement de la population et du logement de 1990 et de

Tableau 7
Totaux estimés selon la taille du ménage pour les personnes de moins de 80 ans en Norvège au 1^{er} janvier 1993 d'après l'estimateur d'Horvitz-Thompson modifié pour l'EDC, l'EEP de 1990 et les projections, en centaines

Taille du ménage		Horvitz-Thompson modifié pour l'EDC		EEP 1990		Projections	
	%		%		%		%
1	622 900	35	626 000	35	668 300	37	
2	518 500	29	494 200	28	549 000	30	
3	259 900	15	291 500	16	221 900	12	
4	258 500	15	250 000	14	221 500	12	
≥ 5	124 600	7	115 300	6	97 500	5	
Inconnue					78 500	4	
Total	1 784 400	1	1 777 000	99	1 826 700	100	

Tableau 8
Probabilité estimée de réponse basée sur la méthode utilisée dans l'EDC depuis 1993, en pourcentage

Taille du ménage		Méthode EDC		Modèle p_{yx} dans (3.1) combiné à RM2(y, z)	
Lieu de résidence					
1	44,53	66,24	74,55	73,54	80,07
Région rurale					
2	36,01	57,90	67,25	66,09	73,80
Région urbaine					
3	47,77	60,90	79,05	73,26	81,52
Région rurale					
4	38,92	52,04	72,44	65,62	75,46
Région urbaine					

L'Enquête sur la famille et la profession de 1988, Keilman et Brunborg calculent des projections pour l'ensemble de la population en 1992. Nous avons ajusté leurs estimations au groupe des 0 à 79 ans.

Les estimations présentées au tableau 7 renforcent notre impression que les estimations fondées sur la modélisation du mécanisme de réponse donne des résultats moins biaisés que ceux obtenus en ne tenant pas compte de ce mécanisme, comme dans la simple poststratification ou la simple expansion. Il en est particulièrement ainsi pour les ménages d'une personne et pour le total. L'« estimateur officiel » courant, c'est-à-dire l'estimateur d'Horvitz-Thompson modifié, semble produire des estimations dont l'ordre de grandeur est correct et qui, en fait, s'approchent plus des résultats de l'EEP de 1990 que les estimations fondées sur un modèle. Cependant, ces résultats sont plus accidentels qu'autre chose. En tant que méthode, l'estimateur modifié présente certains problèmes même dans le cas d'un échantillon représentatif. Nous pouvons les étudier en estimant les probabilités de réponse. Le tableau 8 donne les résultats ainsi que les estimations fondées sur RM2(y, z) et (3.1) tirés du tableau 3.

Tableau 5
Nombres totaux estimés des ménages pour les personnes de moins de 80 ans en Norvège au 1^{er} janvier 1993, exprimés en centaines.
Erreur-type estimée des estimations entre parenthèses

Taille du ménage, y	Estimateur du maximum de vraisemblance avec mécanisme de réponse non ignorable	%	Modèle de population $p_{y,x}$ et modèle de réponse $RM2(y,z)$	%	Modèle de population et de réponse saturé	%	Estimateur ignorable	Mécanisme de réponse
1	558 800	32	595 400	34	596 600	34	486 000	29
2	520 200	30	525 800	30	523 600	30	507 800	30
3	278 900	16	249 100	14	250 000	14	286 200	17
4	258 900	15	269 000	15	268 900	15	270 600	16
≥ 5	125 800	7	126 000	7	126 200	7	131 300	8
Total	1 742 600	100	1 765 300	100	1 765 300	100	1 681 900	100
	(25 600)		(29 700)		(31 900)		(23 300)	
	(4 700)		(5 100)		(5 000)		(4 700)	
	(9 800)		(11 600)		(11 500)		(10 100)	
	(13 800)		(20 300)		(19 800)		(14 100)	
	(38 900)		(48 000)		(53 500)		(35 800)	
	(20 600)		(27 400)		(29 800)		(20 000)	
	(278 900)		(249 100)		(250 000)		(286 200)	
	(258 900)		(269 000)		(268 900)		(270 600)	
	(125 800)		(126 000)		(126 200)		(131 300)	
	(1 742 600)		(1 765 300)		(1 765 300)		(1 681 900)	
	(25 600)		(29 700)		(31 900)		(23 300)	

Tableau 6
Erreurs-types estimées des différences entre les estimations ponctuelles du tableau 5

Taille du ménage		Es1 – Es2	Es1 – Es3	Es2 – Es3	Es4 – Es3
1	29 700	37 000	16 600	42 400	
2	19 300	22 200	8 800	23 100	
3	15 400	15 200	5 300	15 500	
4	6 700	6 500	1 800	6 600	
≥ 5	1 700	1 700	500	1 900	
Total	15 300	18 800	8 900	23 300	

Population : 19,2 – 19,8 – 19,0 – 25,8 – 16,2
Échantillon : 18,6 – 23,0 – 17,8 – 24,9 – 15,7

La proportion, dans l'échantillon, de personnes appartenant à une famille de deux personnes est beaucoup trop élevée et, bien que nous ayons corrigé pour le biais de non-

réponse, l'estimateur par expansion, et donc aussi l'estimateur d'Horvitz-Thompson modifié, ne peut corriger la non-représentativité d'un échantillon. Ceci produira nécessairement des estimations biaisées de H_2 . Nous devons recourir à la poststratification pour corriger l'asymétrie d'un échantillon. Nous pouvons considérer la différence entre les valeurs prévues de ces estimateurs de H_2 comme étant proche du biais de l'estimateur d'Horvitz-Thompson modifié, et nous notons qu'un intervalle de confiance à 95 % approximatif pour cette différence est (39 800, 48 400).

Pour des raisons de robustesse, nous présentons aussi les estimations d'après le modèle logit cumulatif mentionné à la section 3.1 combiné au modèle de réponse $RM1(y,z)$ qui, nous le savons, est mal ajusté aux données. Ces estimations sont exprimées en centaines : 591 800, 501 000, 265 200, 267 300, 128 200 et 1 753 500 pour H_1 , H_2 , H_3 , H_4 , H_5 et H_6 respectivement. Comparativement au tableau 5, ces valeurs

semblent indiquer qu'un modèle raisonnable pour la réponse joue un rôle plus important qu'un bon modèle de population. Il est également évident que la modélisation de la non-réponse importe, comme le montre la comparaison à la poststratification et à l'expansion simple.

5.2 Comparaison avec les estimations utilisées à l'heure actuelle pour l'EDC, l'Enquête sur la qualité du Recensement de 1990 et une étude de projection

Depuis 1993, un estimateur d'Horvitz-Thompson modifié de type (4.14), dont le calcul est plus simple, est utilisé pour produire les statistiques officielles d'après les données de l'EDC (voir Belisby 1995). Nous avons indiqué à la section 2 que les poids sont les inverses des probabilités d'échantillonnage des ménages, multipliés par la probabilité de réponse estimée. Les probabilités de réponse sont estimées au moyen d'un modèle logistique semblable à $RM2(y,z)$ en utilisant le lieu de résidence et la taille du ménage comme variables explicatives. Pour les non-répondants dont on ne connaît pas la taille du ménage, on utilise la taille enregistrée de la famille, en remplaçant (3.5).

Notons que la taille moyenne de la famille pour les familles de 5 personnes ou plus est égale à 670 528/127 653 = 5,25. Nous utilisons 5,25 comme estimation de la taille moyenne du ménage pour les ménages de 5 personnes ou plus et nous divisons par 5,25 au lieu de 5 dans toutes les estimations de H_5 .

5.1 Estimation du maximum de vraisemblance et poststratification

Les distributions estimées des ménages sont présentées au tableau 5. Les estimations sont fondées sur l'estimateur du maximum de vraisemblance (emv) (4.1) en utilisant le modèle de population avec la fonction de lien paramétrique combinée P_{yx} et $RM2(y, z)$. Pour illustrer l'effet de la modélisation de la non-réponse par opposition à celui de la poststratification, nous présentons aussi l'estimateur post-stratifié standard (4.4). Rappelons qu'il s'agit de l'estimateur du maximum de vraisemblance lorsque l'on ne tient pas compte du mécanisme de non-réponse. En outre, nous présentons la distribution estimée de la taille des ménages d'après la poststratification basée sur l'imputation (4.3) avec le modèle saturé (4.9). Pour évaluer la variabilité d'échantillonnage des divers estimateurs, nous incluons aussi les estimations des erreurs-types.

Les trois modèles qui tiennent compte du mécanisme de réponse donnent des nombres totaux de ménages plus élevés. Ils produisent aussi des nombres considérablement plus élevés de ménages d'une seule personne. Ce résultat nous paraît raisonnable, puisque nous attendons à ce que le taux de non-réponse le plus élevé soit celui observé pour cette catégorie de ménages. Et, par conséquent, c'est la prise en compte du mécanisme de non-réponse qui a le plus d'influence sur ces estimations. Nous notons que le modèle à fonction de lien paramétrique contraint (3.1) conjugué au modèle de réponse logarithmique $RM2(y, z)$ donne pratiquement les mêmes estimations poststratifiées que le modèle (4.9), ainsi qu'approximativement les mêmes erreurs-types. Étant donné le niveau de liberté du modèle (4.9), avec un ajustement parfait, il semble que le modèle (3.1) et le mécanisme de réponse $RM2(y, z)$ donnent de bons résultats pour estimer le nombre de ménages de diverses tailles. En ce qui concerne l'incertitude des estimations, nous constatons, comme on pourrait s'y attendre, que les erreurs-types semblent typiquement augmenter avec le nombre de paramètres inconnus dans le modèle sous-jacent. En outre, le nombre total de ménages est estimé de façon assez précise, si l'on ne tient pas compte du biais éventuel, tandis qu'il est manifestement plus difficile d'estimer le nombre de ménages d'une personne. Afin de déterminer dans quelle mesure les différences entre les estimations sont dues à l'erreur d'échantillonnage

échantillonnage aléatoire simple conditionnel, sachant la taille de la famille, la seule hypothèse que nous faisons pour l'estimation de la variance. Inconditionnellement, nous avons un échantillon autopondéré, mais pas aléatoire simple et, par conséquent, il s'agit d'une approximation assez grossière du plan de sondage conditionnel réel. Cependant, pour une étude comparative des estimateurs, l'approximation suffira. L'indicateur de non-réponse r_i est considéré comme étant une constante associée à la personne i . Nous tirons l'échantillon bootstrap, en rééchantillonnant $(y_i, z_i, r_i = 1), (z_i, r_i = 0)$ aléatoirement avec remise, comme il l'est décrit dans Shao et Sitter (1996, section 5), dans chaque poststrate de $\{i; x_i = x\}$. Bien que les tailles des poststrates de l'échantillon soient fixes, aussi bien le nombre de non-répondants que le nombre de personnes provenant d'une région urbaine ou rurale varie d'un échantillon bootstrap à l'autre. Nous calculons les estimations bootstrap de la même façon que celles fondées sur les données observées. En particulier, nous imputons les données bootstrap de la même façon que les données originales si l'estimateur est basé sur l'imputation. Enfin, nous obtenons les variances et les erreurs-types estimées par l'approximation habituelle de Monte Carlo fondée sur 500 échantillons bootstrap indépendants.

5. Nombres estimés de ménages de tailles différentes d'après l'Enquête sur les dépenses de consommation de la Norvège de 1992

Nous présentons ici les nombres estimés de ménages de taille un à cinq et plus, et le nombre total de ménages dans la population norvégienne de moins de 80 ans. L'estimation est faite d'après les données de l'EDC de 1992 et fondée sur les estimateurs décrits à la section 4. Pour calculer les estimations, nous devons connaître les nombres de familles des différentes tailles dans la population, c'est-à-dire M_x , au moment de l'enquête de 1992. Le nombre réel au moment de l'enquête n'a pas été enregistré. À titre d'approximation, nous utilisons les nombres au 1^{er} janvier 1993.

Ils sont donnés au tableau 4.

Tableau 4
Nombre de familles et de personnes de moins de 80 ans en Norvège le 1^{er} janvier 1993

Nombre de personnes dans la famille		Familles	Personnes
1 personne	793 869	793 869	793 869
2 personnes	408 440	816 880	816 880
3 personnes	261 527	784 581	784 581
4 personnes	266 504	1 066 016	1 066 016
5 personnes ou plus	127 653	670 528	670 528
Total	1 857 993	4 131 874	4 131 874

Pour les dix essais multinomiaux déterminés par les différentes valeurs (x, z, z) , nous avons 50 probabilités de cellule inconnues $\pi_{yxz} = P(X = y, R = 1 | x, z)$. En l'absence de contraintes sur les probabilités de cellule, les estimations du maximum de vraisemblance (evm) sont données par les fréquences relatives observées,

$$\hat{\pi}_{yxz} = \frac{n_{xyz}}{n_{xy}} = \frac{x}{z} \frac{m_{xz}}{m_{xy}}.$$

Ce résultat tient aussi quand $u_{xy}(z) = 0$. Maintenant, nous pouvons montrer qu'il existe une correspondance biunivoque entre $(\pi_{xy}^1, \pi_{xy}^2, \dots, \pi_{xy}^s; x; 1, \dots, 5)$ et $(d^1, d^2, \dots, d^s; y; 1, \dots, 5)$: $\pi_{xy}^1 = d^1$, $\pi_{xy}^2 = d^2$, $\pi_{xy}^3 = d^3$, $\pi_{xy}^4 = d^4$, $\pi_{xy}^5 = d^5$, et de d^y doivent satisfaire

$$(4.10) \quad \frac{(z)^{nx}u + (z)^xu}{(z)^{ux}u} = z^{\frac{1}{x}}b, \quad z^{\frac{1}{x}}d$$

et sont déterminées de façon unique par π_{yx}^* . Considérons $h_{xy}(z)$, donné par (4.5) et (4.6). Soit $h_y(z) = \sum^x h_{xy}(z) = \sum^x n_{xy}(z)$ et $n_y(z) = \sum^x n_{xy}(z)$. D'après (4.7),

$$y^f(z) = \frac{h^{f_x}(z) u^f(z) + h^{f_j}(z)}{h^{f_x}(z)}, \text{ si } n_{x_j}(z) > 0. \quad (4.11)$$

De (4.10) et (4.11), il découle que les estimations intuitives suivantes sont aussi des env.

$$(4.12) \quad \frac{u_y(z)u_y(z) + u_y(z)u_y(z)}{u_y(z)} = \hat{q}_{y,z}$$

$$\hat{p}_{z^{\gamma}x^{\gamma}} = \frac{m^x(z) + m^x(z)}{n^{\gamma}(z) + y^{\gamma}(z)} \quad (4.13)$$

(Nous pouvons aussi montrer (4.12) et (4.13) en maximisant directement la log-vraisemblance.) Puis, nous montrons que $h_{xy}(z)$, D'après (4.2), nous avons $h_{xy}^*(z) = m^{xy}(z) \cdot \hat{p}(X = y | x, z, r = 0)$. Sous le modèle (4.9) et les estimations (4.12) et (4.13), nous trouvons que

$$P(Y=x, Z=y | X=z, R=0) = \frac{P(Y=x, Z=y, R=1 | X=z) P(R=1 | X=z)}{P(Y=x, Z=y, R=1 | X=z) + P(Y=x, Z=y, R=0 | X=z)}$$

$$\frac{d^{2\gamma_X} p^{2\gamma_X} - \sum_{\gamma_X} p^{2\gamma_X} \gamma_X}{d^{2\gamma_X} p^{2\gamma_X}} = \frac{(z)^{\gamma_X} u}{(z)^{\gamma_X} y} = \frac{(z)^{\gamma_X} u}{(z)^{\gamma_X} y} =$$

Le modèle (4.9) est saturé et donnera, d'après (4.10), un ajustement parfait.

Les estimations par expansion basée sur l'imputation (4.6), avec le modèle (4.9), sont identiques aux estimations d'Horvitz-Thompson modifiées avec $\hat{q}_{y,c} = n^*(z)/[n^*(2) + n^*(z)]$ (provenant de (4.12)) pour les probabilités estimées de réponse utilisées pour la production des statistiques officielles d'après les données de l'EDC de 1992. Ceci découle du fait que l'estimateur d'Horvitz-Thompson modifié de N_y est donné par

$$\hat{N}_{y, HT} = \sum_{i \in s_i} \frac{I(Y_i = y) \pi_i}{\pi_i},$$

où $\pi_i = P$ (que la personne i soit sélectionnée dans l'échantillon et réponde). Donc,

$$\frac{1}{(\chi = \chi)I} \sum_{i \in s_i} = {}_{\text{HT}} \hat{N}$$

modifié de N_y est donné par

$$\frac{N}{n} = (Y = 1 | X, Z) = R_i \frac{N}{n} = \pi_i$$

$$(4.14) \quad \hat{N}_{\text{IH}}^{\gamma} = \frac{u}{N} \left(\frac{\hat{b}_{\gamma 1}}{(1)_{\gamma u}} + \frac{\hat{b}_{\gamma 0}}{(0)_{\gamma u}} \right)$$

$$\cdot \frac{u}{\delta u + \delta u} N =$$

Donc, cet estimateur d'Horvitz-Thompson modifié souffre de la même caractéristique négative que l'estimateur par expansion basé sur l'imputation (4.6): il ne peut corriger le biais dans un échantillon non représentatif. Pour une description générale de la méthode d'Horvitz-Thompson modifiée, consulter, par exemple, Sæmstad et coll. (1992, chapitre 15).

4.3 Estimation de la variance

Nous estimons la variance des diverses estimations par la méthode du bootstrap. Cette estimation peut se faire sous le cadre de référence de la modélisation ou de la quasi-randémisation (Little et Rubin 1987). Par exemple, pour estimer la variance sous le modèle (3.1) et le mécanisme de réponse RMI (3.2), nous pouvons appliquer le bootstrap paramétrique avec les paramètres estimés (Efron et Tibshirani 1993). Cependant, la façon de comparer les variances estimées sous les divers modèles n'est pas claire. Nous avons, par conséquent, choisi d'estimer les variances des divers estimateurs sous un cadre commun de quasi-randémisation. Nous supposons que nous avons affaire à un

Le résultat général suivant tient, ce qui montre qu'avec le modèle de population (3.1), l'estimateur du maximum de vraisemblance (4.1) est identique à un estimateur poststratifié basé sur l'imputation.

Théorème. Supposons que Y est donné par le modèle (3.1). Autrement dit, $P(Y = y | x, z) = p_{y,x}$ est indépendant de z , mais par ailleurs les $p_{y,x}$ sont complètement inconnues, la seule contrainte étant $\sum_y p_{y,x} = 1$, pour toutes les valeurs de x . Le mécanisme de réponse est paramétrisé arbitrairement, c'est-à-dire qu'aucune hypothèse n'est faite au sujet de $P(R = 1 | Y = y, x, z)$. Alors, les estimations du maximum de vraisemblance pour $p_{y,x}$ sont données, pour $x = 1, \dots, K$, par

$$\hat{p}_{y,x} = \frac{n_{xy} + n_{x*}}{m_x + m_{x*}},$$

où n_{xy} est le nombre de répondants appartenant à une famille de taille x et à un ménage de taille y , $m_x (m_x)$ est le nombre de répondants appartenant à une famille de taille

$$x (z \geq K), \text{ et } m_{x*} = m_{x*}(0) + m_{x*}(1).$$

Preuve. Voir l'annexe A2.

Le théorème implique que l'estimateur peut s'écrire sous forme de l'estimateur poststratifié basé sur l'imputation, en utilisant la taille de la famille comme variable de stratification,

$$\hat{H}_{y, \text{post}}^I = \frac{1}{\sum_x K} \sum_{x=1}^K M_x \frac{n_{xy} + m_{x*}}{m_x + m_{x*}}. \quad (4.3)$$

Si nous supposons que le mécanisme de réponse est ignorable et que nous utilisons le modèle (3.1), la fonction de vraisemblance est donnée par $\prod_{i=1}^n P(Y_i = y_i | x_i)$. Alors, la fonction du maximum de vraisemblance $P(Y = y | x)$ est simplement le taux observé chez les répondants dont la taille du ménage est y , sachant la taille x de la famille. Donc, l'estimateur du maximum de vraisemblance s'avère être identique à l'estimateur poststratifié standard, avec la taille de la famille comme variable de stratification,

$$\hat{H}_{y, \text{post}}^I = \frac{1}{\sum_x K} \sum_{x=1}^K M_x \frac{n_{xy}}{m_x}. \quad (4.4)$$

Pour une étude générale de la poststratification, voir, par exemple, Holt and Smith (1979) et Särndal, Swensson et Wretman (1992, chapitre 7.6).

Afin d'illustrer les effets de la modélisation de la non-réponse et de la poststratification, nous présentons aussi des estimations fondées sur l'estimateur par expansion ordinaire, donné par

$$\hat{H}_{y,e}^I = \frac{1}{N} \cdot N \frac{n_y}{n_x} \quad (4.5)$$

et par l'estimateur par expansion basé sur l'imputation donné par

$$\hat{H}_{y,e}^I = \frac{1}{n_y + n_{x*}} \cdot N \frac{n_y}{n_{x*}}. \quad (4.6)$$

Ici, n_y est le nombre de répondants dans les ménages de taille y , n_{x*} est le nombre total de répondant, et $n_{x*} = \sum_y n_{xy}$. L'estimateur (4.5) ne vise pas à corriger pour la non-réponse ni n'utilise la distribution de la population de familles comme outil de poststratification pour améliorer l'estimation, tandis que l'estimateur (4.6) essaye de tenir compte du mécanisme de non-réponse, mais ne peut fournir une correction pour les échantillons non représentatifs.

4.2 Poststratification basée sur l'imputation avec un modèle saturé

Nous passons maintenant à une méthode intuitive d'imputation qui a été utilisée pour estimer les probabilités de réponse pour un estimateur d'Horvitz-Thompson modifiée pour la production des statistiques officielles d'après les données de l'EDC de 1992 (décrite dans Belsby 1995). Nous utilisons cette méthode d'imputation pour l'estimateur poststratifié (4.3).

La méthode d'imputation consiste à répartir, dans une région rurale/urbaine, les $m_{x*}(z)$ unités non-répondantes sur les ménages de taille $1, \dots, 5$ de telle façon que, sachant la taille du ménage, le taux de non-réponse soit le même pour toutes les tailles de familles. Cette méthode repose sur l'hypothèse implicite que la probabilité de réponse des personnes dont la taille du ménage est la même dans une région rurale/urbaine est identique pour diverses tailles de familles. Soit $h_{xy}(z)$ le nombre de non-répondants dont la taille de la famille est x , la taille du ménage est y et le lieu de résidence est z . Le nombre correspondant de répondants est $n_{xy}(z)$. Les valeurs de $h_{xy}(z)$ sont déterminées au moyen des équations

$$\frac{h_{xy}(z)}{h_{xy}(z)} = \frac{h_{xy}(z) + n_{xy}(z)}{h_{xy}(z)}, \quad z = 0, 1. \quad (4.7)$$

Si $n_{xy}(z) = 0$, nous posons que $h_{xy}(z) = 0$. L'équation (4.7) est résolue sous les conditions

$$\sum_y h_{xy}(z) = m_{x*}(z); x = 1, 2, 3, 4, 5 \text{ et } z = 0, 1. \quad (4.8)$$

La résolution des équations (4.7) et (4.8) nécessite, pour chaque valeur de z , une ligne $(n_{x_1}(z), n_{x_2}(z), \dots, n_{x_5}(z))$ de valeurs non nulles, ce qui est vérifié dans notre cas. Les valeurs imputées $h_{xy}(z)$ déterminées par (4.7) et (4.8) correspondant à la méthode d'imputation décrite par (4.2) pour le modèle suivant :

$$P(Y = y | x, z) = p_{y,x,z} \text{ sans contraintes} \quad (4.9a)$$

$$P(R = 1 | Y = y, x, z) = q_{y,z}, \text{ indépendant de } x. \quad (4.9b)$$

Nous pouvons le montrer comme suit :

Les probabilités de réponse estimées reflètent le taux de réponse plus faible des ménages d'une seule personne et le taux de réponse plus faible dans les régions urbaines. Les ménages de cinq personnes et plus sont ceux dont le taux de réponse est le plus élevé. D'après les modèles, la probabilité estimée de réponse est, curieusement peut-être, plus élevée pour les ménages de trois personnes que pour ceux de quatre personnes. Ce résultat pourrait tenir au fait que les femmes choisissent souvent d'avoir deux enfants et que les ménages de trois personnes sont, pour la plupart, constitués d'une mère, d'un père et d'un *petit* enfant. Ce genre de famille a tendance à rester à la maison et, donc, à être plus accessible qu'une famille de quatre personnes typiques avec deux enfants plus âgés.

Le taux estimé de réponse plus élevé pour les ménages de trois personnes que pour ceux de quatre personnes équivaut à ce que le ratio $P(X = 3 | R = 1) / P(X = 4 | R = 0)$ soit plus grand que le ratio $P(X = 4 | R = 1) / P(X = 3 | R = 0)$. Ceci concorde avec la distribution de la taille des ménages du tableau 2, où nous estimons que $P(X = 4) \approx P(X = 3 | R = 1)$, c'est-à-dire $P(X = 4 | R = 0) \approx P(X = 3 | R = 1)$. Par ailleurs, les estimations du tableau 2 indiquent que $P(X = 3 | R = 1) > P(X = 3)$, ce qui signifie que $P(X = 3 | R = 1) > P(X = 3 | R = 0)$.

Nous voyons que le modèle logistique RM2 combiné au modèle de population avec fonction de lien paramétrique $P_{Y,X}$ contraint à un effet de lissage sur les estimations fondées sur le modèle saturé donné par le modèle (4.9), à cause de l'hypothèse supplémentaire de parallélisme des fonctions logit des probabilités de réponse pour les régions urbaine et rurale.

4. Estimateurs des totaux selon la taille du ménage

À la présente section, nous présentons les estimateurs des totaux selon la taille du ménage, ainsi que la méthode d'estimation de la variance. Nous utilisons un estimateur du maximum de vraisemblance avec la fonction de lien paramétrique contrainte donnée par (3.1) comme modèle de population. Nous montrons que cet estimateur est identique à un estimateur poststratifié basé sur l'imputation, qui de nous, s'avère être un estimateur par poststratification standard si l'on ne tient pas compte du mécanisme de réponse. De surcroît, nous présentons un estimateur poststratifié imputé, basé sur un modèle saturé pour la taille du ménage et la probabilité de réponse.

4.1 Estimateurs fondés sur une fonction de lien paramétrique contrainte comme modèle de population

Si N_y représente le nombre total de personnes vivant dans un ménage de taille y , le nombre de ménages de taille

y est égale à $H_y = N_y / y$. Le nombre total de ménages est représenté par $H, H = \sum_y H_y$. Le problème statistique consiste à estimer H_y pour $y = 1, \dots, J$ et H . Nous choisissons la taille la plus grande J de façon que les ménages de taille supérieure à J soient peu nombreux. Strictement parlant, H_y est le nombre de ménages de taille égale ou supérieure à J , et il en est de même pour N_y . Pour notre application, nous choisissons $J = 5$ car la fréquence des ménages de plus de cinq personnes est faible dans l'échantillon. Nous pouvons écrire $N_y = \sum_{i=1}^J I(X_i = y)$, où la fonction indicatrice $I(X_i = y) = 1$ si $X_i = y$, et 0 autrement. Donc, avec $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$

$$E(H_y | \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N P(X_i = y | \mathbf{x}_i).$$

Nous pouvons obtenir un estimateur par la méthode du maximum de vraisemblance pour H_y en estimant $E(H_y | \mathbf{x})$, c'est-à-dire en remplaçant $P(X_i = y | \mathbf{x}_i)$ par l'estimateur du maximum de vraisemblance $\hat{P}(X_i = y | \mathbf{x}_i)$. Les données sont stratifiées en fonction de la taille de la famille 1, ..., K , où la dernière catégorie contient les personnes appartenant aux familles de taille $\geq K$. Si nous utilisons le modèle avec la fonction de lien paramétrique contrainte définie par (3.1), nous supposons que X dépend uniquement de la taille de la famille x , et l'estimateur prend la forme

$$\hat{H}_y = \frac{1}{J} \sum_{x=1}^K M^x \hat{P}(X = y | x) \quad (4.1)$$

où M^x (M^K) représente le nombre de personnes dans la population dont la taille enregistrée de la famille est x ($\geq K$). Les M^x sont des données auxiliaires connues provenant du Registre norvégien des familles.

Une approche courante pour corriger la non-réponse consiste à imputer les valeurs manquantes dans l'échantillon. En nous fondant sur la distribution estimée de X pour une taille de famille et un lieu de résidence donnés pour les non-répondants, $\hat{P}(X = y | x, z, r = 0)$, nous affectons les non-répondants aux valeurs 1, ..., 5 dans les proportions données par $\hat{P}(X = y | x, z, r = 0)$ pour $y = 1, \dots, 5$. Soit $n_{xy}^*(0)$ le nombre de valeurs imputées pour la taille de famille x et la taille de ménage y , pour les régions rurales (urbaines) et soit $m_{xz}^{rx}(0)$ le nombre d'observations manquantes pour les personnes dans les régions rurales (urbaines) dont la taille de la famille est x . Alors

$$n_{xy}^*(z) = m_{xz}^{rx}(z) \cdot \hat{P}(X = y | x, z, r = 0), \quad z = 0, 1, \quad (4.2)$$

est le nombre total de valeurs imputées pour lesquelles la taille de la famille est x et la taille du ménage est y , c'est-à-dire, n_{xy}^* est le nombre prévu estimé de ménages de taille y , sachant la taille de la famille x et $r = 0$.

Interprétons certaines valeurs obtenues au moyen du modèle fondé sur le ménage. C'est pour les familles à une seule personne que la prise en compte du mécanisme de réponse a l'effet le plus important sur la distribution estimée de la taille du ménage. La probabilité estimative que la taille d'un ménage soit égale à un, sachant que la taille de la famille est de un, est égale à 60,01 %. L'estimation fondée sur l'approche habituelle, en ne tenant pas compte de la non-réponse, est de 51,23 %. Le modèle de réponse « ajuste » le taux observé chez les répondants pour produire une valeur plus élevée, ce qui semble raisonnable, puisque le taux de non-réponse est plus élevé pour les ménages de petite taille. La probabilité estimée que la taille d'un ménage soit de cinq ou plus, sachant que la taille de la famille est égale ou supérieure à cinq, est de 85,55 %, valeur qui diffère peu du taux observé chez les répondants, soit de 87,44 %. Ce résultat indique que, pour une taille de la famille égale ou supérieure à cinq, la distribution de la taille du ménage est à peu près la même chez les répondants que chez les non-répondants.

Le tableau 3 donne les probabilités de réponse estimées en combinant le modèle de réponse RM12 au modèle de population (3.1). En outre, nous présentons les probabilités

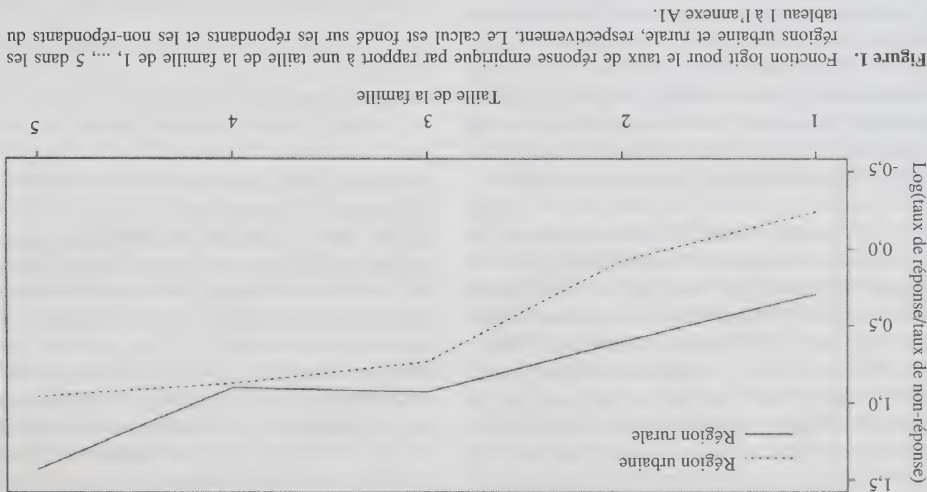


Figure 1. Fonction logit pour le taux de réponse empirique par rapport à une taille de la famille de 1, ..., 5 dans les régions urbaine et rurale, respectivement. Le calcul est fondé sur les répondants et les non-répondants du tableau 1 à l'annexe A1.

Tableau 2 EDC de 1992, Estimation des paramètres, en pourcentage, pour le modèle de population avec fonction de lien paramétrique continue, $p_{x,z}$, combinée au modèle de réponse logistiqu RM12 (y, z). Les estimations pour le modèle de population en ignorant le mécanisme de réponse figure entre parenthèses

Taille de la famille, x		Taille du ménage	
1	2	3	4
60,01 (51,23)	26,75 (29,63)	8,35 (12,35)	4,09 (5,56)
5,27 (3,91)	79,80 (76,98)	12,48 (16,09)	1,47 (1,74)
7,53 (4,72)	14,45 (11,79)	56,67 (61,79)	18,85 (18,87)
1,06 (0,67)	5,31 (4,33)	11,38 (12,33)	77,20 (77,00)
0,84 (0,48)	2,60 (1,93)	1,96 (1,93)	9,05 (8,21)
5 ou plus			85,55 (87,44)

de réponse estimées basées sur un modèle saturé, avec ajustement parfait, décrit à la section 4.2. Le modèle, défini par (4.9), suppose que la probabilité de réponse des personnes faisant partie d'un ménage de même taille dans une région rurale ou urbaine, respectivement, est identique pour différentes tailles de familles. De surcroît, le modèle de la taille du ménage dépend du lieu de résidence et de la taille de la famille, mais sans contraintes sur la fonction de lien. Nous notons que RM12 (y, z) satisfait (4.9b), mais est plus restrictif. Le modèle (4.9) offre plus de liberté que le modèle (3.1) combiné à RM12 (y, z).

Tableau 3 Probabilité estimée de réponse basée sur le modèle logistiqu RM12 combiné à (3.1), et sur le modèle saturé (4.9). Les estimations sont données en pourcentage

Taille du ménage		Taille de la famille	
1	2	3	4
le modèle RM12			
Probabilités de réponse estimées pour			
47,77	60,90	79,16	73,26
38,92	52,04	72,44	65,62
35,17	50,85	74,79	70,68
le modèle saturé			
Probabilités de réponse estimées pour			
50,79	62,37	76,90	70,57
35,17	50,85	74,79	70,68
35,17	50,85	74,79	70,68

Pour $i = 1, \dots, n_i$, L_i est donnée par (3.4) et pour $i = n_i + 1, \dots, n$, L_i est donnée par (3.5).

Nous obtenons les estimations en maximisant la fonction de vraisemblance (3.6). La maximisation a été faite numériquement à l'aide du logiciel TSP (1991) (voir Hall, Cummins et Schnake 1991). L'algorithme d'optimisation est celui de la méthode type du gradient utilisant les dérivées premières et secondes analytiques. Celles-ci sont calculées par le logiciel. L'ajustement du modèle est fondé sur la statistique du chi carré et sur les valeurs de t , fournies par TSP, où les erreurs-types sont calculées d'après les dérivées secondes analytiques. Les valeurs de t doivent être interprétées avec une certaine prudence, puisque l'absence de biais dans les erreurs-types s'écrit dépend de la qualité de la spécification du modèle, ainsi que du nombre d'observations comparativement au nombre de paramètres.

3.3 Évaluation des modèles de la taille du ménage et de la réponse

Nous présentons l'évaluation de l'ajustement des modèles au moyen de la statistique de qualité de l'ajustement de Pearson. L'étude de modélisation est fondée sur les données de l'EDC de 1992. Nous considérons que les paramètres sont significatifs si la valeur absolue de t est supérieure à 2. Cependant, nous ne voulons pas d'un modèle trop restrictif et, par conséquent, nous gardons certaines variables même si leur valeur de t absolue est inférieure à 2.

Dans les modèles de réponse RM1 et RM2, nous utilisons la variable $z = z$, lieu de résidence. Soit $z = 0$ s'il s'agit d'une région rurale et $z = 1$, s'il s'agit d'une région urbaine. On a observé lors de l'EDC de 1986-1988 et de l'EDC de 1992-1994, voir Statistics Norway (1990, 1996), que la non-réponse est plus fréquente durant l'été. Par conséquent nous avons également inclus le moment de l'enquête dans le modèle, c'est-à-dire une variable indiquant si les données ont été recueillies ou non entre le 21 mai et le 12 août. Toutefois, nous avons constaté que le moment de la réalisation de l'enquête n'était pas significatif, la valeur de t étant clairement inférieure à 2. Nous avons également noté que l'effet de la taille de la famille n'était pas significatif. Par contre, si l'on omet la variable de taille du ménage dans le modèle de réponse, alors l'effet de la taille de la famille devient significatif. Idéalement, nous aimerions examiner la fonction logit empirique de la réponse en fonction de la taille du ménage. Cependant, cette dernière est inconnue pour les non-répondants. Par conséquent, nous représentons graphiquement la fonction logit en fonction de la taille de la famille;

voir la figure 1. Si la taille de la famille passe de un à deux, les fonctions pour les familles rurales et urbaines augmentent à peu près parallèlement. Cependant, pour une taille de famille de trois et de quatre, les fonctions logit cessent d'être linéaires et parallèles. Donc, nous pensons que le codage de la taille du ménage sous forme de variable nominale, comme dans le modèle RM2, donnera un meilleur ajustement que la contrainte obligeant les fonctions logit à être parallèles pour les régions rurales et urbaines et linéaires en fonction de la taille du ménage, comme dans le modèle RM1.

Afin de tester la qualité de l'ajustement des modèles, nous considérons la statistique du chi carré de Pearson, sachant les variables auxiliaires x , z . Compte tenu du type rural ou urbain du lieu de résidence et de la taille enregistrée de la famille, il existe six résultats possibles, c'est-à-dire des ménages de taille 1, ..., 5 et la non-réponse. En tout, nous avons dix essais multinomiaux et soixante cellules. Pour les tailles de familles (1,2) et (4,5), les tailles de ménage extrêmes (4,5) et (1,2), respectivement, sont combinées, parce que les tailles attendues sous les modèles sont trop petites. Ceci réduit le nombre de cellules à 52. Le nombre de degrés de liberté (d.d.l.) est calculé comme suit : nombre de cellules - nombre d'essais - moins nombre de paramètres. Pour le modèle (3.1) et RM1 (y , z), d.d.l. = 52 - 10 - (20 + 3) = 19, et pour (3.1) et RM2 (y , z), d.d.l. = 52 - 10 - (20 + 6) = 16. Pour le modèle (3.1) et RM1 (y , z), la statistique χ^2 de Pearson est égale à 26,35 et la valeur p est 0,121. Pour le modèle (3.1) et RM2 (y , z), χ^2 est 21,77 et la valeur p est 0,151.

En étudiant les résidus standardisés, (observé-attendu) / $\sqrt{\text{Var}(\text{observé})}$, nous constatons que la raison principale du meilleur ajustement est que le modèle (3.1) avec RM2 (y , z) prédit mieux les dénombrements observés pour la région urbaine où le taux de réponse est le plus faible (voir l'annexe A1). Donc, les données indiquent que le codage de la taille du ménage sous forme de variable nominale, comme dans RM2, améliore l'ajustement comparativement à l'utilisation d'une variable ordinale. Le modèle (3.1), avec la fonction de lien paramétrique contrainte, combiné à RM2 est le meilleur des modèles que nous avons examinés jusqu'à présent.

3.4 Distribution estimée de la taille du ménage et probabilités de réponse

Le tableau 2 donne les estimations pour le modèle de population (3.1) combiné au modèle de réponse logistique RM2 donné par (3.3).

Ici, α et γ sont des paramètres scalaires et ψ est un vecteur. La variable y_i possède un ordre. Motivé par ce fait, et pour éviter d'introduire de nombreux paramètres, nous utilisons y_i dans (3.2) comme variable ordinale plutôt que comme variable de classe. Donc, la fonction logit,

$$\log\{P(R_i = 1 | y_i, z_i) / P(R_i = 0 | y_i, z_i)\} = \alpha + \gamma y_i + \psi' z_i, \quad (3.3)$$

est linéaire en y_i . Pour éviter l'hypothèse de fonction logit linéaire en y_i , nous considérons également un modèle où

$$RM2(y, z) : P(R_i = 1 | y_i, z_i) =$$

$$\frac{1 + \exp\left(-\alpha_0 - \alpha_1 I_1(y_i) - \alpha_2 I_2(y_i) - \alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi' z_i\right)}{-\alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi' z_i}, \quad (3.3)$$

où la variable indicatrice $I_j(y_i)$ est égale à 1 si $y_i = j$ et 0 autrement. L'inconvénient de ce modèle est qu'il comprend trois paramètres de plus que le modèle (3.2).

3.2 Estimation des paramètres par la méthode du maximum de vraisemblance

Toutes les personnes sélectionnées dans l'échantillon proviennent de ménages différents (les unités d'échantillon-ménage en double ont été éliminées), de sorte que le modèle de population suppose que les tailles de ménage Y_i sont statistiquement indépendantes. Pour cette variable, l'effet d'interaction ou de mise en grappes ne joue aucun rôle.

Considérons la fonction de vraisemblance pour estimer les paramètres inconnus, en supposant que toutes les paires (Y_i, R_i) sont indépendantes et que le modèle de réponse RM1 est donné par (3.2). Pour simplifier la notation, nous ré-annotons les observations de sorte que les observations 1 à n_r soient les répondants et les observations $n_r + 1$ à n soient les non-répondants. Dans le cas du modèle de réponse RM2, l'expression de la vraisemblance est de la même forme avec (3.2) remplacé par (3.3).

$$L_i = \frac{1 + \exp(-\alpha - \gamma y_i - \psi' z_i)}{1} \cdot P_{y_i, x_i}, \quad i = 1, \dots, n_r \quad (3.4)$$

Pour les non-répondants, soit $L_i = P(R_i = 0 | x_i)$. Alors

$$L_i = \sum_{j=1}^y \frac{1 + \exp(\alpha + \gamma y_j + \psi' z_i)}{1} \cdot P_{y_j, x_i}, \quad i = n_r + 1, \dots, n. \quad (3.5)$$

La fonction de vraisemblance pour l'échantillon complet des personnes pour les divers ménages est donnée par

$$L(\theta, \beta, \alpha, \gamma, \psi) = \prod_{i=1}^n L_i. \quad (3.6)$$

Cependant, un test de qualité de l'ajustement, avec x représentant la taille de la famille et le lieu de résidence, indique que ce modèle est mal ajusté aux données. Nous choisissons donc de le rejeter.

Nous supposons que la probabilité de non-réponse peut dépendre de la taille du ménage. Par exemple, les ménages d'une personne sont moins susceptibles de répondre que les ménages de plus grande taille, puisqu'il est plus facile de « trouver quelqu'un à la maison » dans le cas de ces derniers. La non-réponse est indiquée par la variable R_i , où $R_i = 1$ si la personne i répond et 0 autrement. Soit R_i le vecteur de ces indicatrices dans l'échantillon total. D'après Bjørnstad (1996), nous définissons le mécanisme de réponse (MR), c'est-à-dire la loi conditionnelle de R_i sachant les valeurs de x dans la population et les valeurs de y dans l'échantillon total, comme étant ignorable s'il peut être écarté dans une analyse fondée sur la vraisemblance. Cela signifie que le mécanisme de réponse est ignorable si cette loi conditionnelle de R_i ne dépend pas des valeurs observées de y , ce qui coïncide avec la définition utilisée par Little et Rubin (1987, pages 90, 218). Ici, nous supposons que toutes les paires (Y_i, R_i) sont indépendantes. Alors, le mécanisme de réponse MR est ignorable si Y_i et R_i sont indépendants. Par conséquent, le mécanisme de réponse non ignorable est équivalent à

$$P(Y_i = y_i | x_i, r_i = 0) \neq P(Y_i = y_i | x_i, r_i = 1)$$

et, alors, tous deux sont différents de $P(Y_i = y_i | x_i)$.

Donc, estimer les paramètres du modèle pour $P(Y = y | x)$ en utilisant uniquement l'échantillon de répondants et en ne tenant pas compte du fait que la probabilité de réponse dépend de la taille du ménage produirait fort probablement des estimations biaisées des paramètres inconnus. En outre, l'estimateur par poststratification produirait des estimations biaisées, parce qu'il repose sur l'hypothèse que la distribution de R dépend uniquement des variables auxiliaires x . Par exemple, le taux de réponse observé plus faible chez les familles d'une personne indique qu'il pourrait en être de même des ménages d'une personne. Le cas échéant, la probabilité estimée pour un ménage de taille unitaire, fondée uniquement sur les répondants, serait trop faible. La poststratification en fonction de la taille de la famille ne corrigerait fort probablement qu'une partie de ce biais.

Nous supposons que le modèle de la probabilité de réponse, sachant les variables auxiliaires et la taille du ménage y_i , est logistique. Il dépend des variables auxiliaires z_i , qui incluent une partie de x_i , ce qui est exprimé par

$$P(R_i = 1 | y_i, z_i) = \frac{1 + \exp(-\alpha - \gamma y_i - \psi' z_i)}{1}. \quad (3.2)$$

faisons l'analyse statistique sachant l'échantillon total, suivant le principe de vraisemblance (voir Björnsd 1996), fondées sur le plan d'échantillonnage sont sans importance. Il s'agit de l'approche dite de prédiction. Cependant, lorsque nous évaluons l'incertitude statistique des méthodes d'estimation, nous le faisons sous un angle commun de randomisation décrit à la section 4.3.

Pour l'EDC, le vecteur de variables auxiliaires comprend la taille de la famille, le lieu de résidence subdivisé en régions rurales et urbaine et le moment de la collecte des données.

3.1 Les modèles

Considérons d'abord un modèle simple de la taille du ménage, noté Y . Soit x toutes les variables auxiliaires. Nous supposons que la taille du ménage dépend uniquement de la taille de la famille x et, par conséquent, qu'il s'agit d'un modèle avec fonction de lien paramétrique contrainte, mais sans autres hypothèses,

$$P(Y_i = y | x_i) = P(Y_i = y | x_i) = p_{yx_i}, \quad (3.1)$$
$$\sum_y p_{yx_i} = 1, \quad \text{pour chaque valeur possible de } x_i,$$

Le modèle (3.1) est souple en ce sens qu'il ne comprend aucune contrainte sur la fonction hypothétique de x_i . L'inconvénient est le nombre élevé de paramètres comparativement à un modèle de type logistique avec fonction de lien linéaire en x (la fonction reliant $P(Y = y) \mid x$). Si l'on ignore la non-réponse, dans le cas de ce modèle, les estimations seront simplement les taux observés.

La taille du ménage définit des catégories ordonnées. Donc, un choix naturel est le modèle logit cumulé, appelé modèle à odds proportionnels (voir McCullagh et Nelder 1991), en supposant (avec θ_y croissant en y) que

$$P(Y_i \leq y | x) = \frac{1}{1 + \exp(-\theta_y + \beta'x)} \quad \text{pour } y = 1, 2, 3, 4$$

pour $y \geq 5$.

Tableau 1
Tailles de la famille et du ménage pour l'Enquête sur les dépenses de consommation de la Norvège de 1992

Taille du ménage									
Taille de la famille									
1	83	48	20	9	2	3	4	≥ 5	Total
2	9	177	37	4	2	3	17	4	105
3	10	25	131	40	6	17	231	4	267
4	2	13	37	17	17	181	300	17	229
5	1	1	1	1	1	1	1	1	301
6	1	1	1	1	1	1	1	1	209
7	1	1	1	1	1	1	1	1	1 111
8	1	1	1	1	1	1	1	1	587
9	1	1	1	1	1	1	1	1	60
10	1	1	1	1	1	1	1	1	207
11	1	1	1	1	1	1	1	1	123
12	1	1	1	1	1	1	1	1	91
13	1	1	1	1	1	1	1	1	212
14	1	1	1	1	1	1	1	1	6
15	1	1	1	1	1	1	1	1	3
16	1	1	1	1	1	1	1	1	230
17	1	1	1	1	1	1	1	1	160
18	1	1	1	1	1	1	1	1	153
19	1	1	1	1	1	1	1	1	2
20	1	1	1	1	1	1	1	1	162
21	1	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1	1
23	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1	1
29	1	1	1	1	1	1	1	1	1
30	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1
32	1	1	1	1	1	1	1	1	1
33	1	1	1	1	1	1	1	1	1
34	1	1	1	1	1	1	1	1	1
35	1	1	1	1	1	1	1	1	1
36	1	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1	1
40	1	1	1	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1	1
42	1	1	1	1	1	1	1	1	1
43	1	1	1	1	1	1	1	1	1
44	1	1	1	1	1	1	1	1	1
45	1	1	1	1	1	1	1	1	1
46	1	1	1	1	1	1	1	1	1
47	1	1	1	1	1	1	1	1	1
48	1	1	1	1	1	1	1	1	1
49	1	1	1	1	1	1	1	1	1
50	1	1	1	1	1	1	1	1	1
51	1	1	1	1	1	1	1	1	1
52	1	1	1	1	1	1	1	1	1
53	1	1	1	1	1	1	1	1	1
54	1	1	1	1	1	1	1	1	1
55	1	1	1	1	1	1	1	1	1
56	1	1	1	1	1	1	1	1	1
57	1	1	1	1	1	1	1	1	1
58	1	1	1	1	1	1	1	1	1
59	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1
61	1	1	1	1	1	1	1	1	1
62	1	1	1	1	1	1	1	1	1
63	1	1	1	1	1	1	1	1	1
64	1	1	1	1	1	1	1	1	1
65	1	1	1	1	1	1	1	1	1
66	1	1	1	1	1	1	1	1	1
67	1	1	1	1	1	1	1	1	1
68	1	1	1	1	1	1	1	1	1
69	1	1	1	1	1	1	1	1	1
70	1	1	1	1	1	1	1	1	1
71	1	1	1	1	1	1	1	1	1
72	1	1	1	1	1	1	1	1	1
73	1	1	1	1	1	1	1	1	1
74	1	1	1	1	1	1	1	1	1
75	1	1	1	1	1	1	1	1	1
76	1	1	1	1	1	1	1	1	1
77	1	1	1	1	1	1	1	1	1
78	1	1	1	1	1	1	1	1	1
79	1	1	1	1	1	1	1	1	1
80	1	1	1	1	1	1	1	1	1
81	1	1	1	1	1	1	1	1	1
82	1	1	1	1	1	1	1	1	1
83	1	1	1	1	1	1	1	1	1
84	1	1	1	1	1	1	1	1	1
85	1	1	1	1	1	1	1	1	1
86	1	1	1	1	1	1	1	1	1
87	1	1	1	1	1	1	1	1	1
88	1	1	1	1	1	1	1	1	1
89	1	1	1	1	1	1	1	1	1
90	1	1	1	1	1	1	1	1	1
91	1	1	1	1	1	1	1	1	1
92	1	1	1	1	1	1	1	1	1
93	1	1	1	1	1	1	1	1	1
94	1	1	1	1	1	1	1	1	1
95	1	1	1	1	1	1	1	1	1
96	1	1	1	1	1	1	1	1	1
97	1	1	1	1	1	1	1	1	1
98	1	1	1	1	1	1	1	1	1
99	1	1	1	1	1	1	1	1	1
100	1	1	1	1	1	1	1	1	1

Nous considérons un modèle de population hypothétique pour la taille du ménage, sachant les variables auxiliaires, autrement dit, nous modélisons la probabilité conditionnelle. Pour tenir compte de la non-réponse dans l'analyse statistique, nous devons modéliser le mécanisme de réponse, c'est-à-dire la distribution de la réponse sachant la taille du ménage et les variables auxiliaires. Le mécanisme d'échantillonnage des personnes est ignorable pour l'enquête que nous envisageons, autrement dit, est indépendant du vecteur population de tailles du ménage. Par conséquent, nous

3. Modélisation de la taille du ménage et de la non-réponse

Par exemple, le chiffre 48 dans la cellule (1,2) signifie que, des 162 personnes enregistrées comme vivant seules dans l'échantillon répondant, 48 vivent effectivement dans un ménage de deux personnes. Cette situation tient en grande partie au fait que les jeunes gens ont tendance à co-habiter sans être mariés; voir Keilman et Brunborg (1995).

Le tableau 1 montre les données pour l'EDC de 1992 avec un échantillon total de 1 698 personnes. Les ménages dont la taille est égale ou supérieure à cinq sont regroupés à cause de leur faible fréquence dans l'échantillon de ménages. Nous basons notre modélisation et notre estimation sur deux tableaux correspondants, l'un pour les personnes vivant dans les régions rurales et l'autre pour celles vivant dans les régions urbaines. Les données sont présentées au tableau A1 à l'annexe A1.

L'échantillon total, y compris les non-répondants, soit la taille de la famille, le moment de l'enquête (été/pas l'été) et le lieu de résidence (région urbaine/rurale). Les familles sont inscrites dans le Registre norvégien des familles (RNF) et peuvent différer du ménage auquel appartiennent les membres de la famille, par définition ou à cause de changements qui n'ont pas encore été enregistrés. Donc, la taille enregistrée de la famille selon le RNF diffère, dans une certaine mesure, de la taille du ménage. Au départ, d'après l'expérience des enquêtes antérieures, nous supposons que toutes les variables auxiliaires et la taille du ménage influent sur le taux de réponse.

section 3.3, nous évaluons les modèles. Le meilleur ajustement des modèles considérés s'obtient avec un modèle de groupe de tailles de la famille pour la taille du ménage et un lien logistique pour la probabilité de réponse en utilisant la taille du ménage comme variable nominale. À la section 3.4, nous donnons les distributions estimées de la taille du ménage pour diverses tailles de famille et les probabilités de réponses estimées pour diverses tailles de ménages. À la section 4, nous examinons l'estimation fondée sur un modèle, la méthode d'imputation, les estimateurs basés sur l'imputation et la méthode d'estimation de la variance. Nous montrons que, pour le modèle choisi pour la taille du ménage à la section 3.3, l'estimateur du maximum de vraisemblance et l'estimateur poststratifié basé sur l'imputation sont identiques.

À la section 5, nous abordons l'objectif principal, c'est-à-dire l'estimation des nombres totaux de ménages de diverses tailles d'après l'EDC de 1992 en utilisant les estimateurs décrits à la section 4. Le modèle produisant le meilleur ajustement semble donner de bons résultats pour notre problème d'estimation. Nous concluons que la poststratification, la modélisation de la réponse et l'imputation sont des éléments essentiels à l'élaboration d'une méthode satisfaisante.

2. Enquête sur les dépenses de consommation de la Norvège

Les totaux de population selon la catégorie de taille du ménage fournissent des nombres plus corrects de logements que les totaux pour les catégories de taille de la famille déterminées d'après le Registre norvégien des familles. En outre, les autorités chargées de l'évaluation des interventions publiques dans le domaine de la construction de logements se basent sur le nombre estimé de ménages. Par conséquent, estimer les totaux selon la taille du ménage est une question importante en planification sociale. L'EDC est-ignorable, quel que soit le genre d'enquête utilisé, de sorte qu'il s'agit d'une bonne illustration de la façon de traiter les données de l'Enquête sur les dépenses de consommation (EDC) de la Norvège, pour laquelle il est important d'obtenir des renseignements sur la composition des ménages, puisque la taille du ménage influence la consommation.

L'EDC réelle, c'est-à-dire l'enquête sur les variables de dépenses, est réalisée auprès d'un échantillon de ménages privés représentatifs de l'ensemble des ménages privés de la Norvège. Cet échantillon est obtenu en sélectionnant un échantillon de personnes et en incluant tous les membres des ménages auxquels elles appartiennent. Les personnes de plus de 80 ans sont exclues, car elles vivent souvent en

Notre application est fondée sur les données de l'EDC de 1992. L'EDC est une enquête annuelle et, depuis 1992, on utilise un estimateur d'Horvitz-Thompson modifié, comportant une correction pour la non-réponse par estimation des probabilités de réponse sachant la taille du ménage (voir Belsby 1995). Les poids sont égaux à l'inverse de la probabilité de sélection multipliée par la probabilité conditionnelle de réponse sachant que l'unité est sélectionnée. Depuis 1993, la probabilité de réponse est estimée au moyen d'un modèle logistique dont les variables auxiliaires sont le lieu de résidence (région rurale/urbaine) et la taille du ménage. Pour la plupart des non-répondants, on utilise la taille de la famille comme substitut de la taille du ménage. Par exemple, nous entendons les personnes ayant un logement commun et partageant au moins un repas par jour (c'est-à-dire logeant sous le même toit). Pour une description complète de l'EDC, consulter Statistiques Norway (1996). Dans l'EDC, les variables auxiliaires connues pour

établissement. Aux fins de l'étude, les unités d'intérêt de l'enquête sont les *personnes* de 16 à 80 ans vivant dans les ménages privés et la variable d'intérêt est la taille du ménage auquel la personne appartient, qui est observée uniquement dans l'échantillon de personnes sélectionnées répondantes.

Le plan de sondage est un plan d'échantillonnage de personnes aupondéré à trois degrés. Autrement dit, chaque personne faisant partie de la population a la même probabilité d'inclusion dans l'échantillon total. Aux deux premiers degrés, des régions géographiques sont sélectionnées de façon stratifiée, tandis qu'au troisième, des personnes sont sélectionnées aléatoirement à partir des régions géographiques sélectionnées. Au premier degré, les unités primaires d'échantillonnage (UP) sont les municipalités de la Norvège. Celles comptant moins de 3 000 habitants sont regroupées, de sorte que chaque UP soit constituée d'au moins 3 000 personnes. Les UP sont d'abord regroupées en dix régions, puis, dans chaque région, elles sont stratifiées d'après la taille (nombre d'habitants) et le type de municipalité (c'est-à-dire, structure industrielle et centralité). En tout, nous avons obtenu 102 strates. Les villes de plus de 30 000 habitants représentent leur propre strate et, par conséquent, sont sélectionnées avec certitude au premier degré. Pour les autres strates, une UP est sélectionnée avec probabilité proportionnelle à la taille. Au deuxième degré, les UP sont sélectionnées de façon que l'échantillon total résultant de personnes soit représentatif de la population totale. Les UP sont sélectionnées, la taille de l'échantillon est déterminée de façon que l'échantillon total résultant de personnes soit

autopondéré.

Méthodes de modélisation et d'estimation de la taille du ménage en présence de non-réponse non ignorable appliquées à l'Enquête sur les dépenses de consommation de la Norvège

Liv Belsby, Jan Bjørnstad et Li-Chun Zhang¹

Résumé

Nous considérons le problème de l'estimation, en présence de non-réponse non ignorable importante, du nombre de ménages privés de diverses tailles et du nombre total de ménages en Norvège. L'approche est fondée sur un modèle de population pour la taille du ménage enregistré de la famille. Nous évaluons compte du biais de non-réponse éventuel en modélisant le mécanisme de réponse sachant la taille du ménage. Nous évaluons divers modèles, ainsi qu'un estimateur du maximum de vraisemblance et une poststratification fondée sur l'imputation. Nous comparons les résultats à ceux d'une poststratification pure avec la taille enregistrée de la famille comme variable de stratification et des méthodes d'estimation employées pour la production de statistiques officielles d'après l'Enquête sur les dépenses de consommation de la Norvège. L'étude indique que la modélisation de la réponse, la poststratification et l'imputation sont des éléments importants d'une approche satisfaisante.

Mots clés : Taille du ménage; non-réponse; imputation; poststratification.

1. Introduction

La présente étude a été motivée par le taux élevé de non-réponse à l'Enquête sur les dépenses de consommation (EDC) de la Norvège réalisée auprès des ménages privés, qui était de 32 % pour l'enquête de 1992. La non-réponse comprend les impossibilités de prendre contact et les refus de participer. Nous nous concentrons sur le problème de la non-réponse non ignorable qui se pose lorsqu'on estime le nombre de ménages de diverses tailles et le nombre total de ménages.

Nous considérerons une approche entièrement fondée sur un modèle, à savoir la modélisation et l'estimation de la distribution de la taille des ménages sachant la taille enregistrée de la famille et du mécanisme de réponse sachant la taille du ménage. Ce modèle tient compte du fait que le mécanisme de non-réponse pourrait être non ignorable, en ce sens que la probabilité de réponse dépend de la taille du ménage. Le modèle de réponse sert à corriger pour la non-réponse. Les approches fondées sur un modèle avec non-réponse incluse, parfois appelées approches prédictives, ont été considérées, entre autres, par Little (1982), Greenlees, Reece et Zieschang (1982), Baker et Laird (1988), Bjørnstad et Walsøe (1991), Bjørnstad et Skjold (1992), ainsi que Forster et Smith (1998).

Pour divers modèles de taille du ménage et de réponse, nous examinons principalement deux approches fondées sur un modèle, c'est-à-dire un estimateur du maximum de vraisemblance et la poststratification basée sur l'imputation

en fonction de la taille enregistrée de la famille. Nous comparons ces méthodes à la poststratification pure et aux méthodes utilisées à l'heure actuelle pour l'EDC.

La grande question ici est de comparer des modèles et des méthodes dont le problème principal est le biais de non-réponse. En outre, nous estimons par une méthode bootstrap les erreurs-types des estimations et des différences entre les estimations, sachant les tailles des strates a posteriori déterminées d'après la taille de la famille. En plus d'évaluer l'incertitude statistique des estimateurs, ceci nous permet de déterminer dans quelle mesure les différences entre les estimateurs proposés sont attribuables à l'erreur d'échantillonnage, au biais de non-réponse ou aux deux. Cependant, dans cette évaluation, nous gardons à l'esprit la citation qui suit, tirée de Little et Rubin (1987, page 67) : [traduction] « Il importe d'insister sur le fait que, dans de nombreuses applications, le problème du biais de non-réponse est souvent plus crucial que celui de la variance. En fait, d'aucuns ont soutenu que fournir une estimation valide de la variance d'échantillonnage est pire que ne pas fournir d'estimation si l'estimateur présente un biais important, qui domine l'erreur quadratique moyenne. »

À la section 2, nous décrivons la structure des données et le plan de sondage de l'EDC, et à la section 3, nous examinons les problèmes de modélisation. À la section 3.1, nous présentons les divers modèles de taille du ménage et de réponse à prendre en considération pour l'EDC de 1992, à la section 3.2, nous décrivons la méthode du maximum de vraisemblance pour l'estimation des paramètres et à la

1. Liv Belsby, Statistique Norvège, Division des méthodes statistiques et des normes, C.P. 8131 Dep., N-0033 Oslo, Courriel : lbe@ssb.no; Jan F. Bjørnstad, Statistique Norvège, Division des méthodes statistiques et des normes, C.P. 8131 Dep., N-0033 Oslo, Courriel : jab@ssb.no; Li-Chun Zhang, Statistique Norvège, Division des méthodes statistiques et des normes, C.P. 8131 Dep., N-0033 Oslo, Courriel : lcz@ssb.no.

Bibliographie

- Ciuff, A.D., et Ord, J.K. (1981). *Spatial Processes, Models and Applications*. Piton Limited, London.
- Cox, D.R., et Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society*, Series B, 49, 1-39 (avec discussion).
- Cressie, N. (1990). Small-Area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, October 1990.
- Cressie, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquêtes*, 18, 83-103.
- Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimation problem. *Statistica Sinica*, 10, 613-627.
- Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistics Association*, 74, 267-277.
- Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
- Kackar, R.N., et Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistics Association*, 79, 853-862.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1997). Rapport No. 404(50/1.0/1). Consumption of some important commodities in India. NSS 50th Round, juillet 1993-juin 1994.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 436(51/1.0/1). Household consumption expenditure and employment situation in India. NSS 51st Round, juillet 1994-juin 1995.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 440(52/1.0/1). Household consumption expenditure and employment situation in India. NSS 52nd Round, juillet 1995-juin 1996.
- Ciuff, A.D., et Ord, J.K. (1981). *Spatial Processes, Models and Applications*. Piton Limited, London.
- Cox, D.R., et Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society*, Series B, 49, 1-39 (avec discussion).
- Cressie, N. (1990). Small-Area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, October 1990.
- Cressie, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquêtes*, 18, 83-103.
- Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimation problem. *Statistica Sinica*, 10, 613-627.
- Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistics Association*, 74, 267-277.
- Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
- Kackar, R.N., et Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistics Association*, 79, 853-862.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1997). Rapport No. 404(50/1.0/1). Consumption of some important commodities in India. NSS 50th Round, juillet 1993-juin 1994.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 436(51/1.0/1). Household consumption expenditure and employment situation in India. NSS 51st Round, juillet 1994-juin 1995.
- National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 440(52/1.0/1). Household consumption expenditure and employment situation in India. NSS 52nd Round, juillet 1995-juin 1996.
- Shrivastava, V.K., et Tiwari, R. (1976). Evaluation of expectations of products of stochastic matrices. *Scandinavian Journal of Statistics*, 3, 135-138.
- Singh, A.C., Sukei, D. et Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 377-396.
- Singh, A.C., Manel, H.J. et Thomas B.W. (1994). MPLSE à données chronologiques pour petites régions évaluées à l'aide de données d'enquête. *Techniques d'enquêtes*, 20, 35-46.
- C. Spall). New York: Marcel Dekker Inc., 477-508.
- Singh, A.C., Manel, H.J. et Thomas B.W. (1994). MPLSE à données chronologiques pour petites régions évaluées à l'aide de données d'enquête. *Techniques d'enquêtes*, 20, 35-46.
- Singh, A.C., Sukei, D. et Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 377-396.
- Srivastava, V.K., et Tiwari, R. (1976). Evaluation of expectations of products of stochastic matrices. *Scandinavian Journal of Statistics*, 3, 135-138.
- Ripley, B.D. (1981). *Spatial Statistics*. New York: John Wiley & Sons, Inc.
- Sallas W.M., et Harville D.A. (1994). Noninformative priors and restricted likelihood estimation in the Kalman filter. *Bayesian Analysis of Time Series and Dynamic Models*, (Ed. James C. Spall). New York: Marcel Dekker Inc., 477-508.
- Singh, A.C., Manel, H.J. et Thomas B.W. (1994). MPLSE à données chronologiques pour petites régions évaluées à l'aide de données d'enquête. *Techniques d'enquêtes*, 20, 35-46.
- Singh, A.C., Sukei, D. et Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 377-396.
- Srivastava, V.K., et Tiwari, R. (1976). Evaluation of expectations of products of stochastic matrices. *Scandinavian Journal of Statistics*, 3, 135-138.

$$S^{\wedge}_p(\psi) = \text{Col}[S^p_a(\psi)], \quad S^p_a(\psi) = \frac{\partial \psi^p}{\partial \epsilon}.$$

En utilisant l'expression pour les dérivées de la vraisemblance, nous obtenons

$$S^p_a(\psi) = \frac{1}{2} \left[\text{Trace}[C^p_a(\psi)] - \sum_{i=1}^I \text{Trace} \left[H^{-1}_{i-1} \frac{\partial \psi^p}{\partial H_i} \right] + \sum_{i=1}^I [\epsilon'_i B^p_{ii}(\psi) \epsilon_i] \right]$$

$$- \left[\epsilon'_i X^T_{i-1} H^{-1}_{i-1} \frac{\partial \psi^p}{\partial \epsilon_i} C^p_a(\psi) \right] = \left[(X^T_1 H^{-1}_1 X^{-1}_1)^T X^T_1 H^{-1}_1 \frac{\partial \psi^p}{\partial H_1} H^{-1}_1 X_1 \right],$$

$$B_{ii}^p(\psi) = H^{-1}_{i-1} \frac{\partial \psi^p}{\partial H_i} H^{-1}_i.$$

En tenant compte de ce que $\epsilon'_i \sim N(0, H_i)$, $\text{Corr}(\epsilon'_i, \epsilon_j) = 0$ pour $i \neq j$, $\text{Corr}(\epsilon'_i, (\partial \psi^p / \partial \psi^p)) = 0$ en raison du fait que $\text{Corr}(\epsilon'_i, (\partial \psi^p / \partial \psi^p)) = 0$ en raison du fait que $(\partial \epsilon'_i / \partial \psi^p) = (\partial (y_i - U^T \alpha^{m-1})) / (\partial \psi^p)$ est une fonction linéaire de $(y_1, y_2, \dots, y_{i-1})$ et n'est donc pas corrélée à ϵ_i , nous obtenons l'espérance des termes les plus intérieurs de l'expression (5.13) sous la forme

$$[E[\epsilon'_i S^p_a(\psi) S^p_e(\psi) E^T_i]] = K^{ae}_p(\psi) H_i + 2 \left[\frac{\partial \psi^p}{\partial H_i} H^{-1}_{i-1} \frac{\partial \psi^p}{\partial H_i} \right] \left[\frac{1}{2} \text{Trace}[B_{ii}^p(\psi)] + \text{Trace}[B_{ie}^p(\psi)] \frac{\partial \psi^p}{\partial H_i} \right] + \frac{1}{4} \text{Trace}[B_{ii}^p(\psi)] \text{Trace}[B_{ie}^p(\psi)] H_i,$$

où

$$K^{ae}_p(\psi) = \frac{1}{I} \sum_{i=1}^I \text{Trace} \left[H^{-1}_{i-1} \frac{\partial \psi^p}{\partial H_i} H^{-1}_{i-1} \frac{\partial \psi^p}{\partial H_i} \right].$$

Les trois termes médians de l'expression étant d'ordre $O(1)$, ce qui, conjugué à $I^{-1}(\psi)$ dans l'expression donnée ci-dessous, les rend d'ordre $o(m^{-1})$,

$$E[\langle \theta_i, (\psi) - \theta_i, (\psi) \rangle \langle \theta_i, (\psi) - \theta_i, (\psi) \rangle^T] = g_{3i}(\psi) = L^T_x [(\psi) \otimes I^{-1}_{i-1} (\psi) \otimes I^m] [K^{\wedge}(\psi) \otimes H]$$

$$[I^{\wedge}_{i-1} (\psi) \otimes I^m] [L(\psi) \otimes I^m + o(m^{-1})]$$

$$= L^T_x [(\psi) \otimes I^{-1}_{i-1} (\psi) K^{\wedge}(\psi) \otimes I^{-1}_{i-1} (\psi) \otimes L(\psi) \otimes I^m] + o(m^{-1}).$$

$$\|\psi - \psi\|.$$

L'ordre $o(m^{-1})$ (Peers et Iqbal 1985).

Le deuxième terme de l'expression $[(\psi - \psi)^T \otimes I^m \Delta^2 g_{12i}(\psi)]$ est la variance asymptotique de ψ . Le premier terme de l'expression $[(\psi - \psi)^T \otimes I^m \Delta^2 g_{12i}(\psi)]$ se réduit à $g_{4i}(\psi)$, parce que $E[(\psi - \psi) = b^{\wedge}_i(\psi)]$ jusqu'à

$$\begin{aligned} &= -g_{5i}(\psi) = g_{5i}(\psi) \\ &= \left[\frac{1}{2} \text{Trace} \left[\left(\frac{\partial \psi^p}{\partial \Sigma} \right)^T \left(\frac{\partial \psi^p}{\partial \Sigma} \right) \right] \otimes (\Sigma^{-1} R) \right] \\ &= -L^T_x [(\psi) \otimes I^m \Delta^2 g_{12i}(\psi)] [(\psi) \otimes I^m] \end{aligned}$$

pression sous la forme

symétriques, nous obtenons le deuxième terme de l'ex-

Compte tenu du fait que $\Sigma(\psi)$ et ses dérivées sont

$$\begin{aligned} \frac{\partial^2 g_{12i}(\psi)}{\partial \Sigma^{-1} \partial \Sigma^{-1}} \frac{\partial \psi^p}{\partial \Sigma^{-1}} \frac{\partial \psi^p}{\partial \Sigma^{-1}} R &= -2R \Sigma^{-1} \frac{\partial \psi^p}{\partial \Sigma^{-1}} \frac{\partial \psi^p}{\partial \Sigma^{-1}} \frac{\partial \psi^p}{\partial \Sigma^{-1}} R \\ &+ R \Sigma^{-1} \frac{\partial \psi^p}{\partial \Sigma^{-1}} \frac{\partial \psi^p}{\partial \Sigma^{-1}} R \\ \Delta^2 g_{12i}(\psi) &= \text{Col}[\Delta^2 g_{12i}(\psi)] = \text{Concat} \left[\frac{\partial^2 g_{12i}(\psi)}{\partial \psi^p \partial \psi^p} \right]_{1 \leq p \leq 3} \\ &= R \Sigma^{-1} \frac{\partial \psi^p}{\partial R} \frac{\partial \psi^p}{\partial \Sigma^{-1}} R \end{aligned}$$

$$\Delta^2 g_{12i}(\psi) = \text{Col}[\Delta^2 g_{12i}(\psi)] = \text{Concat} \left[\frac{\partial^2 g_{12i}(\psi)}{\partial \psi^p \partial \psi^p} \right]_{1 \leq p \leq 3}$$

$$+ o^p(m^{-1})$$

$$+ \frac{1}{I} [(\psi - \psi)^T \otimes I^m \Delta^2 g_{12i}(\psi)] [(\psi - \psi) \otimes I^m]$$

$$g_{12i}(\psi) = g_{12i}(\psi) + [(\psi - \psi) \otimes I^m]^T \Delta^2 g_{12i}(\psi)$$

$g_{12i}(\psi)$ autour de ψ , nous obtenons

La preuve est essentiellement du même type que celle du théorème A.2. Par développement en série de Taylor de

Preuve du théorème A.4

$$E[g_{5i}(\psi)] = g_{5i}(\psi) + o(m^{-1}).$$

et

$$E[g_{3i}(\psi)] = g_{3i}(\psi) + o(m^{-1})$$

$$= g_{12i}(\psi) + o(m^{-1})$$

$$E[g_{12i}(\psi) + g_{3i}(\psi) + g_{31i}(\psi) - g_{4i}(\psi) - g_{5i}(\psi)]$$

Théorème A.4 : Sous les conditions de régularité 2

et de $(\partial^2 \theta(\psi)) / (\partial \psi^p \partial \psi^q) |_{\psi=\psi_0} = O^p(1)$ quand $\|\psi^* - \psi\|$ nous donne

$$[\hat{\theta}(\psi) - \theta(\psi)] = [\psi - \psi] \otimes I^m \nabla \hat{\theta}(\psi) + O^p(m^{-1}). \quad (5.2)$$

!c, $\nabla \hat{\theta}(\psi) = (\partial \hat{\theta}(\psi)) / (\partial \psi) = [(\partial \hat{\theta}(\psi)) / (\partial \psi)]$.

$(\partial \psi^2)^T$. En utilisant

$$\frac{\partial \psi^p}{\partial \psi} = \sum_p \frac{\partial \psi^p}{\partial \psi} \frac{\partial \psi^p}{\partial \psi} \Big|_{\beta=\beta(\psi)} + \frac{\partial \psi^p}{\partial \psi} \frac{\partial \psi^p}{\partial \psi} \Big|_{\beta=\beta(\psi)}$$

$$p = 1, 2$$

où $\hat{\theta}^*(\beta, \psi) = X\beta(\psi) + A(\psi)[\gamma - X\beta(\psi)]$, et le fait que $(\partial \hat{\theta}^a(\psi)) / (\partial \psi^p) = O^p(m^{-1/2})$ (Cox et Reid 1987), nous

trions de ce qui précède

$$[\hat{\theta}(\psi) - \theta(\psi)] = [\psi - \psi]^T \otimes I^m \nabla \hat{\theta}^*(\psi) + O^p(m^{-1}). \quad (5.3)$$

$$\text{où } \nabla \hat{\theta}^*(\psi) = \left[\frac{\partial}{\partial \psi} \hat{\theta}^*(\beta, \psi), \frac{\partial}{\partial \psi} \hat{\theta}^*(\beta, \psi) \right]^T \Big|_{\beta=\beta(\psi)}$$

En utilisant les conditions de régularité I et le fait que $\beta(\psi) - \beta = O^p(m^{-1/2})$, nous avons

$$[\hat{\theta}(\psi) - \theta(\psi)]$$

$$= [\psi - \psi] \otimes I^m \nabla \hat{\theta}(\psi) + O^p(m^{-1})$$

$$= \sum_{\gamma=1}^p (\psi^p - \psi^p) L^p(\psi) [\gamma - X\beta(\psi)] + O^p(m^{-1}).$$

En outre, si nous utilisons le développement en série de Taylor de la vraisemblance $S(\hat{\eta}) = 0$ autour de ψ , où

$$S(\eta) = [S_T^{\beta}(\eta), S_T^{\alpha}(\eta)]^T, S_T^{\beta}(\eta) = \text{Col} \left[\frac{\partial}{\partial \eta} \right]_{\eta=\eta_0}^p$$

et l'orthogonalité de β et ψ , il s'ensuit que

$$(\psi - \psi) = I_{-1}^{\psi} S(\eta) + O^p(m^{-1}).$$

En écrivant

$$S^{\psi}(\psi) = \text{Col} [S^p(\psi)] = [S^p(\psi), S^{\alpha p}(\psi)]^T,$$

$$S^p(\psi) = \frac{\partial \psi^p}{\partial \eta} = \frac{1}{\text{Trace}} \left[\frac{\partial \psi^p}{\partial \eta} \right] + \frac{1}{2} [u^T B^p(\psi) u],$$

$$B^p(\psi) = \Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \text{ et } \gamma = \gamma - X\beta(\psi)$$

$$I^p(\psi) = \frac{1}{\text{Trace}} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right]$$

nous obtenons

jusqu'à l'ordre $o(m^{-1})$

En introduisant cette expression par substitution dans (5.4) et étant donné que la deuxième expression est d'ordre $O(m^{-1})$, nous pouvons obtenir l'expression suivante

$$E[us^p(\psi) S^p(\psi) u^T] = \frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right] + \frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right].$$

(1976), elle devient

et, en appliquant les résultats de Srivastawa et Tiwari

$$E[us^p(\psi) S^p(\psi) u^T] = \begin{bmatrix} u[u^T B^p(\psi) u] u^T [u^T B^p(\psi) u] u^T & -u \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right] [u^T B^p(\psi) u] u^T \\ -u \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right] [u^T B^p(\psi) u] u^T & -n \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right] [u^T B^p(\psi) u] u^T \end{bmatrix}$$

l'intérieur dans l'expression (5.4) devient

L'espérance d'un élément type des termes les plus à

où la matrice d'information $I^{\psi}(\psi) \equiv I^{\psi}(\psi)$.

$$E \left[-\frac{\partial^2 \ell}{\partial \psi^p \partial \psi^q} \right] = \frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \Sigma^{-1} \frac{\partial \psi^q}{\partial \eta} \right] = I^{pq}(\psi)$$

$$B^p(\psi) = \Sigma^{-1} \frac{\partial \psi^p}{\partial \eta}$$

$$\frac{\partial \ell}{\partial \psi^p} = -\frac{1}{2} \text{Trace} \left[\Sigma^{-1} \frac{\partial \psi^p}{\partial \eta} \right] + \frac{1}{2} u^T B^p(\psi) u,$$

$$\ell = \log L = \text{const.} - \frac{1}{2} \log |\Sigma| - \frac{1}{2} u^T \Sigma^{-1} u$$

dérivée sous la forme

Maintenant, nous pouvons écrire la vraisemblance et sa

$$(5.4) \quad I^{\psi}(\psi) \otimes I^m [L(\psi)]$$

$$= L^T(\psi) [I^{\psi}(\psi) \otimes I^m] \text{Col} [\text{Concat} [us^p(\psi) S^p(\psi) u^T]]$$

$$I^{\psi}(\psi) \otimes I^m [L(\psi)]$$

$$= L^T(\psi) [I^{\psi}(\psi) \otimes I^m] \text{Col} [us^p(\psi) S^p(\psi) u^T]$$

$$[\hat{\theta}(\psi) - \theta(\psi)]^T [\hat{\theta}(\psi) - \theta(\psi)]^T \text{ jusqu'à l'ordre } o(m^{-1})$$

et, donc, l'expression

$$[\hat{\theta}(\psi) - \theta(\psi)] = L^T(\psi) [I^{\psi}(\psi) \otimes I^m] [S^{\psi}(\psi) \otimes u]$$

ou « Temp » dénote le modèle spatio-temporel et k_i le modèle 2, 3 ou 4. De même, nous avons déterminé l'efficacité relative du modèle de régression temporel (modèle 4) par rapport au modèle de régression simple (modèle 2) en simulant des données au moyen des estimations des paramètres du tableau 3, sous le modèle de régression temporel. Les résultats sont présentés au tableau 5.

Pour ces paramètres, les résultats confirment la supériorité du modèle spatio-temporel comparativement aux autres. Le modèle de régression temporel s'avère également meilleur que le modèle de régression simple.

5. Conclusion

L'utilisation d'un modèle pour petits domaines, caractéristique du domaine, améliore considérablement les estimations directes par sondage, car elle permet d'exploiter l'autocorrélation spatiale entre les domaines voisins. Cependant, il ne faut appliquer le modèle qu'après avoir déterminé si la corrélation entre les petits domaines en vertu de leurs effets de voisinage est significative. Si la relation entre la variable dépendante et les variables exogènes est faible, le modèle spatial simple avec ordonnée à l'origine seulement peut améliorer tout autant les estimations. Ce modèle ne tire parti que de l'autocorrélation spatiale pour renforcer les estimations par petits domaines et ne requiert pas l'utilisation de variables exogènes. Les modèles spatiaux, grâce à l'utilisation de la matrice W appropriée, ou d'une combinaison de matrices W , peuvent améliorer considérablement les estimations. La matrice de poids devrait s'appuyer sur des considérations logiques et est parfois utile dans les cas où, pour certaines raisons, on ne dispose pas de variables exogènes fiables. Cet aspect peut aussi être exploité pour obtenir les estimations par petits domaines pour les domaines qui ont été créés/délimités récemment.

Il faut faire attention à l'accroissement de l'erreur quadratique moyenne (EQM) causé par la variabilité due au remplacement des paramètres par leurs estimations. Cet aspect, que reflète l'approximation de deuxième ordre de l'EQM examinée dans le présent article, est la raison pour laquelle, très souvent, le simple modèle spatial (avec ordonnée à l'origine) donne de meilleurs résultats que le modèle spatial comprenant un plus grand nombre de paramètres. L'utilisation de données chronologiques avec des paramètres de régression constants au cours du temps améliore encore davantage les estimations par petits domaines, particulièrement pour les points dans le temps où la EQM des estimations directes par sondage est grande. Les modèles spatio-temporels présentent des avantages par rapport aux modèles temporels qui ne tiennent pas compte des effets spatiaux, grâce à l'inclusion d'une autocorrélation

spatiale constante entre les petits domaines. Cependant, pour certains points dans le temps pour lesquels n peut être fort différent de celle des autres points, cet avantage n'est pas nécessairement vérifié, parce que les estimations tendent vers la moyenne des cinq cycles. Dans ce cas, on peut choisir un premier point dans le temps approprié pour commencer à tenir compte des effets temporels. Enfin, les variables exogènes X et la matrice de poids W se complètent par la voie du paramètre de régression β et du paramètre d'autocorrélation ρ , et l'utilisation judicieuse de ces paramètres peut donner lieu à une amélioration importante des estimations par petits domaines.

Remerciements

Les données de niveau unitaire utilisées pour l'étude nous ont été fournies par la National Sample Survey Organisation (NSSO), du ministère des Statistiques et de la Mise en œuvre des programmes aux termes d'une entente de recherche entre IIT Kanpur et la NSSO. La matrice de poids contenant la longueur de la frontière commune pour diverses paires de petits domaines (districts) a été fournie par le centre national d'information (NIC) du ministère des Technologies de l'information du gouvernement de l'Inde. Nous tenons à remercier les examinateurs de leurs commentaires constructifs qui nous ont permis d'améliorer beaucoup l'article.

Annexe

Théorème A.1 : Sous les conditions de régularité 1

$$EQM[\theta(\psi)] = g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (5.1)$$

Pour prouver le théorème, nous utilisons les résultats suivants bien connus (Shrivastawa et Tiwari 1976). Soit $U \sim N(0, \Sigma)$, alors pour les matrices symétriques A , B et C

$$\begin{aligned} E[U(U^T A U) U^T] &= \text{Trace}(A \Sigma) \Sigma + 2 A \Sigma \\ E[U(U^T A U)(U^T B U) U^T] &= \text{Trace}(A \Sigma) \text{Trace}(B \Sigma) \\ &+ 2[\text{Trace}(A \Sigma) \Sigma B \Sigma + \text{Trace}(B \Sigma) \Sigma A \Sigma + \text{Trace}(A \Sigma B \Sigma) \Sigma] \\ &+ 4[\Sigma A \Sigma B \Sigma + \Sigma B \Sigma A \Sigma]. \end{aligned}$$

Preuve du théorème A.1

Kackar et Harville (1984) ont montré que $EQM[\theta(\psi)] = E[(\theta(\psi) - \hat{\theta}(\psi))(\theta(\psi) - \hat{\theta}(\psi))^T]$. Il est facile de montrer que $EQM[\theta(\psi)] = g_1(\psi) + g_2(\psi)$. Nous devons prouver que $g_3(\psi) = E[(\theta(\psi) - \hat{\theta}(\psi))(\theta(\psi) - \hat{\theta}(\psi))^T] + o(m^{-1})$. Le développement en série de Taylor de $\hat{\theta}(\psi)$ autour de ψ et l'utilisation de $(\psi - \psi) = O_p(m^{-1/2})$

Afin d'évaluer les propriétés des estimateurs sous divers (modèle spatio-temporel), nous avons simulé des données sous le modèle spatio-temporel et avons calculé les EQM réelles des estimations répétées sous chacun des modèles étudiés. Pour ce faire, nous avons exécuté la simulation en prenant les paramètres estimés d'après le modèle spatio-temporel présente au tableau 2 et avons calculé la moyenne réelle de petit domaine répétée $\theta(b)$ pour la b^e répétition ($b = 1, 2, \dots, B$) ainsi que les observations simulées $y(b)$ pour un grand nombre de répétitions. Sur cet ensemble de données simulées, pour chaque répétition, nous avons appliqué divers modèles, y compte le modèle spatio-temporel et avons calculé les estimateurs de la moyenne de petit domaine sous chacun.

Lors de l'ajustement du modèle de régression temporel et du modèle spatio-temporel aux ensembles de données simulés, nous avons exécuté le processus de maximisation

$$EQM(\theta_i^j) = \frac{1}{B} \sum_{b=1}^B [\theta_i^j(b) - \theta_i(b)]^2, \quad i = 1, 2, \dots, m.$$

Nous avons évalué l'efficacité relative des estimateurs sous le modèle spatio-temporel (modèle 5) comparative-ment aux estimateurs sous les modèles 2 à 4 au moyen du ratio de leurs erreurs quadratiques moyennes (EQMR) donné par

$$EQMR(k, Temp) = 100 \frac{\sum_{i=1}^m EQM(\theta_i^k)}{\sum_{i=1}^m EQM(\theta_i^{Temp})}$$

Tableau 4

MPLNBE moyens pour les DCMH (R), leur e-r. estimée et leur c.v. sous les modèles de régression, spatial, de régression temporel et spatio-temporel

Modèle	50	51	52	53	54	55
Cycle de la NSSO						
Modèle 1	276,10	321,26	373,07	408,52	411,25	482,00
Modèle 2	272,87	312,53	354,45	397,52	400,87	471,99
Modèle 3	272,98	313,14	351,51	398,21	400,78	471,09
Modèle 3A	273,56	314,19	352,01	396,40	399,91	471,91
Modèle 4	274,13	305,62	345,54	383,53	399,56	463,32
Modèle 5	273,75	312,21	351,79	391,61	399,50	473,57
Erreurs-types moyennes (e-t.)						
Modèle 1	25,09	66,06	64,18	74,19	53,87	45,45
Modèle 2	17,10	33,65	29,09	39,85	32,68	30,59
Modèle 3	16,88	32,84	21,51	39,98	30,87	24,84
Modèle 3A	16,56	31,29	20,79	40,03	30,23	24,37
Modèle 4	19,51	34,91	35,19	37,79	35,14	33,15
Modèle 5	17,18	28,99	28,33	30,02	28,76	28,10
Coefficients de variation moyens (c.v.) (%)						
Modèle 1	9,09	20,56	17,20	18,16	13,10	9,43
Modèle 2	6,27	10,79	8,21	10,01	8,15	6,48
Modèle 3	6,18	10,49	6,12	10,04	7,70	5,27
Modèle 3A	6,05	9,96	5,91	10,10	7,56	5,17
Modèle 4	7,12	11,42	10,18	9,85	8,79	7,15
Modèle 5	6,28	9,29	8,05	7,67	7,20	5,93

Tableau 5

Efficacité relative [EQMR] en pourcentage des modèles temporels comparativement aux autres modèles pour les DCMH

Cycle de la NSSO						
50	51	52	53	54	55	
Modèle spatio-temporel [Modèle 5]						
Modèle 2	123,63	170,54	193,68	203,55	204,72	169,76
Modèle 3	100,24	133,82	149,70	165,46	165,85	154,23
Modèle 4	125,81	141,50	141,93	137,55	139,11	129,88
Modèle de régression temporel [Modèle 4]						
Modèle 2	100,71	134,50	156,35	165,30	163,13	152,56

Lorsque nous comparons le modèle de régression simple (modèle 2) et le modèle spatial (modèle 3) au moyen de la statistique LRT, nous constatons que, sous $H_0(p=0)$, pour le modèle 3, l'autocorrélation spatiale p est hautement significative pour les cycles 52 et 55; de toute évidence, du modèle spatio-temporel.

Dans le cas du modèle 4, le processus de maximisation itérative sans contrainte a convergé pour une valeur de k supérieure à 1, ce qui est inadmissible sous l'hypothèse de stationnarité. Pour ce cas, nous avons obtenu les estimations en prenant $k=1$ et le modèle 4 a été modifié en conséquence. Le tableau 3 donne les résultats pour $k=1$ dans le cas du modèle de régression temporel. Le modèle spatio-temporel produit une valeur plus élevée du coefficient d'autocorrélation commun et une valeur nettement

Le tableau 3 donne les estimations des paramètres et leur erreur-type dans le cas du modèle de régression temporel et du modèle spatio-temporel.

Dans le cas du modèle 4, le processus de maximisation itérative sans contrainte a convergé pour une valeur de k supérieure à 1, ce qui est inadmissible sous l'hypothèse de stationnarité. Pour ce cas, nous avons obtenu les estimations en prenant $k=1$ et le modèle 4 a été modifié en conséquence. Le tableau 3 donne les résultats pour $k=1$ dans le cas du modèle de régression temporel. Le modèle spatio-temporel produit une valeur plus élevée du coefficient d'autocorrélation commun et une valeur nettement

Les estimations directes par sondage sont moins précises et tous les modèles comprenant des termes d'effets mixtes les améliorent. Les estimations pour les cycles 50 et 55 (fondées sur de grands échantillons) sont plus précises que celles obtenues pour les autres cycles. Le modèle spatial, qui dépend de la valeur de p , améliore considérablement les estimations. Dans le cas des cycles 52 et 55, où l'auto-corrélation s'avère significative, la réduction de l'e-r. l. moyen des estimations est importante comparativement aux modèles sans autocorrélation spatiale. Le modèle 3A avec effet spatial et sans variables auxiliaires est tout aussi bon. Le modèle spatio-temporel améliore encore davantage les estimations en tirant parti des considérations d'espace d'états. Il convient de souligner que, pour le cycle 52 (très forte autocorrélation spatiale), les estimations au moyen des modèles temporels sont moins bonnes que celles ne tenant pas compte du temps. Peut-être parce que les paramètres de régression et d'autocorrélation sont fixes, les estimations tendent vers la moyenne des cinq cycles.

Nous résumons les résultats du tableau 4 ci-après.

plus faible de l'estimation de σ_v^2 . Au tableau 4, nous résumons les estimations moyennes par cycle des DCMH (fondées sur l'ensemble des 63 districts), leurs erreurs-types (e-r.) estimées et le coefficient de variation (c.v.) sous chaque modèle.

Tableau 2 Estimations des paramètres pour les estimations par petits domaines des DCMH sous le modèle de régression et les modèles spatiaux

Cycle	R^2	σ_v^2	p	σ_v^2	λ_1	p	σ_v^2	λ_2	LRT
50	0.27	1 724.48	0.30	1 635.70	1.80	0.59	1 724.68	6.64	LRT
51	0.27	(356.19)	(0.18)	(346.45)	(0.13)	0.67	(378.66)	6.64	Modèle 3A
52	0.17	2 150.54	0.87	(815.24)	13.46	(0.13)	(824.54)	4.54	Modèle 3A
53	0.13	6 312.99	-0.39	(257.15)	1.56	(0.07)	(272.27)	7.66	Modèle 3A
54	0.22	(1 397.92)	(0.27)	(1 374.70)	1.30	0.66	(1 561.72)	3.00	Modèle 3A
55	0.31	2 989.73	0.87	1 060.21	20.30	0.86	1 186.58	1.56	Modèle 3A
λ_1 et λ_2 comparant les modèles 2, 3 et les modèles 3, 3A respectivement. $\chi^2_{3,05}=7.815$ pour λ_2 . $\chi^2_{1,05}=3.841$ pour λ_1 et									

Tableau 3 Estimation des paramètres de l'estimation par petits domaines des DCMH sous le modèle de régression temporel et le modèle spatio-temporel

Modèle	Estimation	E.T.	Modèle 4	Estimation	E.T.	Modèle 5	Estimation	E.T.
σ_v^2	4 715.64	2 163.50	431.00	245.50	—	—	—	—
p	—	0.04	—	0.79	—	—	—	—
k	—	—	—	—	—	—	—	—

exogènes ont été sélectionnées par analyse des covariances parmi une foule de variables allant des variables du Recensement de 1991 à celles couvertes par les données annuelles sur l'agriculture. Nous avons examiné diverses matrices de poids, comme la longueur de la frontière commune entre deux districts, la distance entre les centres de district ou les poids binaires. Comme ces derniers donnent une estimation plus grande du coefficient d'auto-corrélation spatiale, nous les avons utilisés pour poursuivre l'analyse présentée ici (après les avoir normalisés en rendant la somme des éléments de chaque ligne de la matrice des poids égale à un). Dans tout l'exercice, nous avons procédé à la maximisation de la fonction de log-vraisemblance et à l'estimation des paramètres par la méthode simple de Nelder et Mead au moyen du logiciel MATLAB.

Divers modèles à effets mixtes utilisés pour trouver des estimations améliorées des DCMH sont présentés au tableau 1. Les paramètres de ces modèles ont la signification habituelle mentionnée aux sections 2 et 3. En outre, dans le cas de chaque modèle, nous supposons que la variance d'échantillonnage R ou R_i (dans le cas du modèle temporel) est connue.

Tableau 1
Modèles à effets mixtes

Modèle - 1	Estimations directes	$y = X\beta + v + e$
Modèle - 2	Modèle de régression	$y = X\beta + Zv + e$
Modèle - 3A	Modèle spatial (ordonnées à l'origine)	$y = \mu + Zv + e$
Modèle - 4	Modèle de régression temporel	$y_t = X_t'\beta + v_t + e_t, v_t = kv_{t-1} + \eta_t$
Modèle - 5	Modèle spatio-temporel	$y_{jt} = X_{jt}'\beta + Zv_{jt} + e_{jt}, v_{jt} = kv_{j,t-1} + \eta_{jt}$

Le tableau 2 présente les estimations du modèle de régression simple et des modèles spatiaux à effets mixtes. La valeur du coefficient de corrélation multiple R^2 entre les estimations des DCMH et les variables auxiliaires est également présentée pour chaque cycle. Les erreurs-types (e.-t.) des estimations des paramètres sont indiquées entre parenthèses. Notons que λ_1, λ_2 est la statistique du test du rapport des vraisemblances (LRT pour *likelihood ratio test*) défini comme étant $-2 \log L \sim \chi^2_k$, où L est le ratio entre les vraisemblances emboîtées et les valeurs hypothétiques des paramètres pour deux modèles concurrents sous diverses hypothèses, et k est la différence entre le nombre de paramètres sous les deux modèles. Ici, λ_1 compare le modèle de régression et le modèle spatial, sous $H_0: \rho = 0$ contre $H_1: \rho \neq 0$ et suit la loi de χ^2_1 sous H_0 , et λ_2 compare le modèle spatial et le modèle spatial (avec ordonnée à l'origine), sous $H_0: \beta = 0$ contre $H_1: \beta \neq 0$ [il n'inclut pas le terme d'ordonnée à l'origine β_0] et suit la loi de χ^2_3 sous H_0 .

annuelle, l'enquête est réalisée auprès d'un petit échantillon comptant quatre ménages par unité de premier degré, tandis que dans la série quinquennale, l'enquête est réalisée auprès d'un échantillon de 10 à 12 ménages par unité de premier degré. À part cela, les enquêtes de la NSSO comportent deux échantillons, c'est-à-dire un échantillon central réalisé par les chercheurs de la NSSO et un échantillon d'Etat effectué par les autorités de l'Etat. En ce qui concerne la procédure d'estimation, les unités de premier degré sont sélectionnées sous forme de deux sous-échantillons indépendants. L'estimation de la moyenne de population et de sa variance est calculée séparément d'après les deux sous-échantillons. La moyenne groupée $y_i = (y_{i1} + y_{i2})/2$ et $R_i = (y_{i1} - y_{i2})^2/4$ pour $i = 1, 2, \dots, m$, où y_{i1}, y_{i2} sont les moyennes de sous-échantillon, estimées respectivement, la moyenne de population et sa variance pour un district (petit domaine) particulier. Dans le cas du cycle 55, les unités de premier degré ont été sélectionnées sous la forme de huit sous-échantillons indépendants et l'estimation de la moyenne de population et celle de sa variance ont été calculées d'après ces sous-échantillons. Étant donné les problèmes que posent les estimations des R_i avec un degré de liberté, le R_i pour chaque petit domaine a été analysé et comparé au cours du temps. Toute valeur anormale de R_i a été lissée par calcul de la moyenne des R_i sur les points dans le temps voisins et, dans certains cas, sur les petits domaines avoisinants également. Les estimations par sondage y_i sont les estimations directes et les R_i lissés sont les éléments diagonaux de la matrice des variances-covariances d'échantillonnage R dans nos équations mo-

Nous n'utilisons ici que les données provenant de l'échantillon central. Nous avons calculé les estimations des dépenses de consommation mensuelles par habitant (DCMH) et des erreurs-types (e.-t.) des estimateurs sous domaines mixtes pour les 63 districts (petits domaines) ruraux d'un grand Etat de l'Inde, à savoir l'Uttar Pradesh. Nous avons utilisé les données provenant de six cycles de la NSSO, c'est-à-dire le cycle 50 (juillet 1993 à juin 1994), le cycle 51 (juillet 1994 à juin 1995), le cycle 52 (juillet 1995 à juin 1996), le cycle 53 (janvier à décembre 1997), le cycle 54 (janvier à juin 1998) et le cycle 55 (juillet 1999 à juin 2000). De ceux-ci, les cycles 50 et 55 sont fondés sur des enquêtes quinquennales. Les variables exogènes choisies pour être utilisées dans les modèles sont i) le nombre de ménages, ii) la superficie brute (la superficie nette ensemencée et iii) la superficie nette ensemencée dans les districts. Les données agricoles sont disponibles sur une base annuelle, tandis que les estimations des ménages et de la population ont été obtenues par une méthode d'interpolation d'après les données des recensement décennaux de 1971, 1981 et 1991. Ces variables

L'approximation de deuxième ordre de l'EQM du

MPLNBE est

$$EQM[\hat{\theta}_i(\psi)] = E[(\hat{\theta}_i(\psi) - \theta_i)(\hat{\theta}_i(\psi) - \theta_i)^T]$$

$$= g_{12i}(\psi) + g_{3i}(\psi) + o(m^{-1}). \quad (3.10)$$

Ici, $g_{3i}(\psi)$ est le biais dû à l'estimation des paramètres d'après des données d'échantillon qui est d'ordre $O(m^{-1})$ et est donné par

$$g_{3i}(\psi) = L_i^T(\psi) I^{-1}(\psi) K^{\psi}(\psi) H_i^T I^{-1}(\psi) L_i(\psi) \quad (3.11)$$

où $K^{\psi}(\psi) \equiv (K^{de}(\psi))$

$$\text{et } K^{de}(\psi) = \frac{1}{T} \sum_{i=1}^T \text{Trace} \left[H_i^{-1} \frac{\partial \psi^p}{\partial H_i} H_i^{-1} \frac{\partial \psi^e}{\partial H_i} \right]. \quad (3.12)$$

En outre,

$$L_i(\psi) = \text{Col} [L_{1i}^{ps1}(\psi)] \text{ et } L_{1i}^{ps1}(\psi) = (\partial \Delta^i(\psi)) / (\partial \psi^p)$$

pour $d = 1, 2, 3$.

Sous une forme correcte, nous pouvons écrire $g_{3i}(\psi)$

comme suit

$$g_{3i}(\psi) =$$

$$\left[\sum_{d=1}^p \sum_{e=1}^p L_{de}^{ps1}(\psi) \times \sum_{f=1}^f \text{Trace} \left(H_i^{-1} \frac{\partial \psi^f}{\partial H_i} H_i^{-1} \frac{\partial \psi^g}{\partial H_i} \right) L_{fg}^{ps1}(\psi) \right] \times H_i^T I_{fg}^{ps1}(\psi)$$

L'expression pour la matrice d'information intervenant ici peut être donnée par

$$I^{de}(\psi) = E \left[- \frac{\partial^2 \psi^p \partial \psi^e}{\partial^2 l} \right] = \frac{1}{T} \sum_{i=1}^T \text{Trace} \left[H_i^{-1} \frac{\partial \psi^p}{\partial H_i} H_i^{-1} \frac{\partial \psi^e}{\partial H_i} \right] + \sum_{i=1}^T \left[\frac{\partial \psi^p}{\partial H_i} H_i^{-1} \frac{\partial \psi^e}{\partial H_i} \right] \times \left[\frac{\partial \psi^p}{\partial H_i} H_i^{-1} \frac{\partial \psi^e}{\partial H_i} \right] \times \left[\frac{\partial \psi^p}{\partial H_i} H_i^{-1} \frac{\partial \psi^e}{\partial H_i} \right] \times \left[\frac{\partial \psi^p}{\partial H_i} H_i^{-1} \frac{\partial \psi^e}{\partial H_i} \right]$$

Nous avons également obtenu l'estimateur de l'EQM du MPLNBE sous l'hypothèse d'une grande valeur de m et en négligeant tous les termes d'ordre $o(m^{-1})$ dans le théorème A.4 en annexe sous la forme

$$\text{eqm}[\hat{\theta}_i(\psi)] = [g_{12i}(\psi) + g_{3i}(\psi) + g_{31i}(\psi) - g_{4i}(\psi) - g_{5i}(\psi)] + o(m^{-1}), \quad (3.13)$$

où $g_{31i}(\psi)$, $g_{4i}(\psi)$ et $g_{5i}(\psi)$ sont donnés par

$$g_{31i}(\psi) = L_i^T(\psi) I^{-1}(\psi) \otimes H_i^T(\psi) L_i(\psi), \quad (3.14)$$

$$g_{4i}(\psi) = [b_i^T(\psi) \otimes I_m] \frac{\partial}{\partial g_{12i}(\psi)}, \quad (3.15)$$

$$g_{5i}(\psi) = \frac{1}{T} \text{Trace} \left[L_i^T(\psi) \otimes (R_i^T H_i^{-1}) \right] \frac{\partial \psi^p \partial \psi^e}{\partial^2 H_i^T} [I^{-1}(\psi) \otimes (H_i^{-1} R_i)], \quad (3.16)$$

4. Analyse des données de la NSSO

La National Sample Survey Organisation (NSSO) du ministère de la Statistique et de la Mise en œuvre des programmes (gouvernement de l'Inde) réalise de grandes enquêtes par sondage quinquennales (EQ) sur les dépenses des ménages et l'emploi, presque tous les cinq ans en Inde. Le champ d'observation de ces enquêtes compte plus de 100 000 ménages répartis entre plusieurs villages et foyers urbains. Afin de combler les lacunes annuelles en matière de consommation (EDC) annuelle enquête sur les dépenses de consommation à une période de six mois ou d'un an). La série annuelle ne couvre que de 10 000 à 30 000 ménages selon le nombre de villages et de foyers urbains visés par l'enquête dans l'ensemble du pays. Chaque cycle de la NSSO porte généralement sur plus d'un thème principal différent. Cependant, le questionnaire I.0 de ces enquêtes est conçu en vue de recueillir des données sur les dépenses de consommation des ménages parmi d'autres caractéristiques de l'emploi.

La NSSO utilise un plan d'échantillonnage stratifié à deux degrés, où les unités de premier degré sont les villages de recensement dans le secteur rural sélectionnés par échantillonnage systématique avec des points de départ aléatoires indépendants. L'Inde est subdivisée en États et les districts sont les unités administratives de deuxième degré dans les États. Les enquêtes annuelles et quinquennales diffèrent peu, à l'exception du fait que, normalement, dans la série

Les estimateurs des effets fixes et aléatoires et de la EOM de ces estimateurs sont obtenus par clape, en partant de l'hypothèse d'une approche par modèle linéaire à effets mixtes au temps $t = 1$, et en prenant $v_1 \sim N^m(0, \sigma_v^2 I)$ (Salas et Harville 1994). Sous la forme standard, nous écrivons le modèle comme suit

$$[\,{}^mI_z^{\alpha,d}0]\mathfrak{diag}=\bar{O}\quad(\bar{O},0)^{m+d}N\sim{}^t\mathfrak{Z}$$

$$(\mathfrak{E})\quad[\,{}^mK_z^{\alpha,d}I]\mathfrak{diag}=T,{}^t\mathfrak{Z}+{}^{t-1}\alpha\mathcal{L}={}^t\alpha,{}^t\mathfrak{Z}+{}^t\alpha{}^tU={}^t\mathcal{U}$$

$$(3.4) \quad U^i = [X^i, Z], \alpha^i = [\beta^i, v^i]^i.$$

Ici, I_m et 0_m sont la matrice unité et la matrice nulle de dimension m , et $\text{diag}[I^p, KI^m]$ représente la matrice

$$\begin{bmatrix} w \times w & d \times w \\ w \times d & d \times d \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & I \end{bmatrix}$$

Si l'on suppose que β est fixe, mais qu'il dépend du temps, le modèle ne change pas, sauf que $T = \text{diag}[0^p, kI^m]$.

Les estimations initiales des effets α_i et de leur variance (basées sur $t = 1$) sont obtenues comme suit

$$\begin{bmatrix} \Sigma_{22} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{11} \end{bmatrix} = \Sigma^{-1} A^{-1} R_1 \Sigma^A A^{-1} H$$

Les équations des filtres de Kalman récurrentes pour la mise à jour des estimateurs aux étapes subséquentes sont

$$\begin{aligned} & {}^{[-1:n]} \overline{\mathbf{z}}' \mathbf{A} {}^{[-1:n]} \mathbf{H} {}^{[-1:n]} \overline{\mathbf{z}} - {}^{[-1:n]} \overline{\mathbf{z}} = {}' \overline{\mathbf{z}} \\ & ({}^{[-1:n]} \mathbf{p}' \mathbf{A} - {}^{[-1:n]} \mathbf{H} {}^{[-1:n]} \overline{\mathbf{z}}) {}^{[-1:n]} \mathbf{H} {}^{[-1:n]} \overline{\mathbf{z}} + {}^{[-1:n]} \mathbf{p} = {}' \mathbf{p} \\ & {}^{[-1:n]} \mathbf{A} {}^{[-1:n]} \overline{\mathbf{z}}' \mathbf{A} + {}^{[-1:n]} \mathbf{y} = {}' \mathbf{H} \quad {}^{[-1:n]} \mathbf{H} \mathbf{A} = {}^{[-1:n]} \mathbf{p}' \overline{\mathbf{z}} + {}^{[-1:n]} \mathbf{z} \mathbf{A} = {}^{[-1:n]} \mathbf{z} \end{aligned}$$

où les α_i sont les estimateurs des effets α_i , sachant les observations $[y_1, y_2, \dots, y_{i-1}]$ et les $\Sigma^{i|i-1}$

A l'aide des équations de y_i , sachant $[y_1, y_2, \dots, y_{j-1}]$, les moyennes conditionnelles de y_j , représentant la matrice covariante des variables latentes, nous obtenons le meilleur prédicteur (MPLNB) de $\theta_i = X_i\beta + Z_i\gamma_i$, et l'erreur quadratique moyenne (EQM) du MPLNB comme suit

$$(3.5) \quad \begin{aligned} & (\mathfrak{h})^i \mathfrak{e}(\mathfrak{h})^i \mathbf{V} + (\mathfrak{h})^i \mathfrak{r}^i \mathfrak{g}(\mathfrak{h})^i \mathbf{I} = \\ & [(\mathfrak{h})^i \mathfrak{r}^i \mathfrak{g}(\mathfrak{h})^i \mathbf{I} - {}^i \chi](\mathfrak{h})^i \mathbf{H}^i \mathbf{R} - {}^i \chi = \\ & (\mathfrak{h})^i \mathfrak{g}(\mathfrak{h})^i \mathbf{I} = (\mathfrak{h})^i \Theta \end{aligned}$$

quadratique moyenne (EQM) du MPLNB comme suit

Soulignons que $\delta_{12}(\psi)$ est l'analogue spatial de $\delta_1(\psi) + \delta_2(\psi)$. Comme il est fréquent en pratique, le vecteur de paramètres ψ est inconnu et ses estimateurs du maximum de vraisemblance restreint (EMVR) peuvent être obtenus en maximisant la fonction de log-vraisemblance suivante, d'après les données d'échantillon couvrant tous les points dans le temps.

$$[H]_{\text{eff}} = \sum_{l=1}^L \frac{1}{l} - [X^T H X]_{\text{eff}} \frac{1}{l} - \text{const} = l$$

$$(L\varepsilon) \quad ({}^{1-n} \vartheta^i \Omega - {}^i \kappa) \cdot {}^1 H \cdot ({}^{1-n} \vartheta^i \Omega - {}^i \kappa) \sum_{i=1}^{z-1} \frac{z}{i} -$$

$$({}^1 \vartheta^i X - {}^i \kappa) \cdot {}^1 H \cdot ({}^1 \vartheta^i X - {}^i \kappa) \frac{z}{i} -$$

en ce qui concerne le paramètre ψ . À l'aide de l'équation qui précède, nous obtenons l'estimateur $\hat{\psi}$, et le MPLNBE de θ , et l'estimateur naïf de l'EOM du MPLNBE sont donnés par

$$\theta(\psi)' \theta(\psi)' U = (\psi)' \alpha(\psi)' U = (\psi)' \alpha(\psi)' \alpha(\psi)^{-1} V + (\psi)' e(\psi)' \psi = (\psi)' V + (\psi)' \Sigma(\psi)' U_T' (\psi)' \cdot \quad (3.9)$$

Comme nous l'avons expliqué plus haut à la section 2, l'EQM du MPLNBE sous-estime l'EQM réelle car elle ne tient pas compte de la variabilité due au remplacement des paramètres par leurs estimations. Au théorème A.3 en particulier, nous obtenons une approximation de deuxième ordre de $E^*[EQM\theta^*(\psi)]$ pour une grande valeur de m et en négligeant tous les termes d'ordre $o(m^{-1})$, sous les conditions de régularité qui suivent satisfaites par notre modèle. Ces conditions sont analogues aux conditions de régularité 1.

Conditions de régularité 2

a) Les éléments de $X^i, i=1, 2, \dots, r$ sont bornés uniformément de sorte que

$$\begin{aligned} \Gamma &= \Gamma, \text{wscut}[(\Gamma)O] = ({}^p\mathfrak{h}e^p\mathfrak{h}e)/([(\mathfrak{h})^jV_z^j e]), \text{dscut}[(\Gamma)O] \\ &= ({}^p\mathfrak{h}e)/([(\mathfrak{h})^j\Lambda(\mathfrak{h})^jV]e), \text{dscut}[(\Gamma)O] = (\mathfrak{h})^j\Lambda(\mathfrak{h})^jV \end{aligned} \quad (c)$$

(d) ψ est l'estimation de ψ de ψ qui satisfait à $\psi(y) = \psi(-y)$, $\psi(y) = \psi(x+y)$ et $\psi(y) = \psi(hy)$ pour tout $y \in R^p$.

est la matrice d'information et \otimes représente le produit de Kronecker. En outre, $g_3(\psi)$ peut aussi s'écrire

$$g_3(\psi) = \sum_{j=1}^p \sum_{l=1}^p L_{jl}(\psi) \Sigma(\psi) L_l^c(\psi) I_{-1}^{dp}(\psi) \quad (2.14)$$

$$\text{ou } I_{-1}^{\psi} \equiv (I_{-1}^{dp}(\psi)).$$

En pratique, il est courant d'estimer l'EQM du MPLNBE par conséquent, nous avons les expressions

$$E[g_1(\psi) + g_3(\psi) - g_4(\psi) - g_5(\psi)] = g_1(\psi) + o(m^{-1}), \quad (2.15)$$

Par conséquent, nous avons les expressions

$$E[g_2(\psi)] = g_2(\psi) + o(m^{-1}) \quad \text{et} \quad E[g_3(\psi)] = g_3(\psi) + o(m^{-1}), \quad (2.16)$$

et, enfin, l'estimateur de l'EQM de $\theta(\psi)$ sous la forme

$$[g_1(\psi) + g_2(\psi) + 2g_3(\psi) - g_4(\psi) - g_5(\psi)] + o(m^{-1}), \quad (2.17)$$

où $E[\text{eqm}(\theta(\psi))] = \text{EQM}[\theta(\psi)] + o(m^{-1})$.

De toute évidence, les termes supplémentaires, $g_3(\psi)$, $g_4(\psi)$ et $g_5(\psi)$ sont les contributions dues à l'estimation du vecteur de paramètres inconnu ψ au moyen de ψ . Les expressions pour $g_4(\psi)$ et $g_5(\psi)$ jusqu'à l'ordre $o(m^{-1})$ sont données par

$$g_4(\psi) = [b_T^{\psi}(\psi) \otimes I_m] \frac{\partial \psi}{\partial g_1(\psi)},$$

$$b^{\psi}(\psi) = \frac{1}{I} \text{Col} \left[\text{Trace} \left[I_{-1}^{\psi}(\psi) \text{Col}^{\top} \left[\frac{\partial I_{\beta}^{\psi}(\psi)}{\partial \psi^p} \right] \right] \right] \quad (2.18)$$

$$g_5(\psi) = \frac{1}{I} \text{Trace} \left[\frac{\partial^2 \psi \psi^{\top}}{\partial^2 \Sigma(\psi)} [I_{-1}^{\psi}(\psi) \otimes \Sigma^{-1}(\psi) R] \right] \quad (2.19)$$

Ici, $b^{\psi}(\psi)$ est le biais de ψ , c'est-à-dire $E(\psi) - \psi$ jusqu'à l'ordre $o(m^{-1})$ et $(\partial g_1(\psi))/(\partial \psi)$, $(\partial g_2(\psi))/(\partial \psi^p)$ de dimension $(2m \times m)$ ayant deux matrices de dimension $m \times m$ dans une colonne. De la même façon, $(\partial^2 \Sigma(\psi))/(\partial \psi \partial \psi^{\top})$ est une matrice partitionnée de dimension $(2m \times 2m)$ ayant deux colonnes, en ce qui concerne les lignes et les colonnes, avec à l'initieur, $(\partial^2 \Sigma(\psi))/(\partial \psi^p \partial \psi^p)$, une sous-matrice générale de dimensions $m \times m$. $\text{Trace}(B) = \sum_{i=1}^m B_{ii}$, où B est une matrice carrée partitionnée en sous-matrices carrées

3. Modèle spatio-temporel

L'expression (2.17) donne la matrice de l'estimateur de l'EQM du MPLNBE, $\hat{\theta}(\psi)$ et l'EQM des estimateurs sur petit domaine individuels est donnée par les éléments diagonaux respectifs. Nous pouvons obtenir des expressions similaires pour un modèle simple sans autocorrélation spatiale. Cependant, dans ce cas, $g_5(\psi)$ devient nulle.

$$g_5(\psi) = \frac{1}{I} \sum_{j=2}^p \sum_{l=2}^p \left[R \Sigma^{-1}(\psi) \frac{\partial \psi^p \partial \psi^c}{\partial^2 \Sigma(\psi)} \Sigma^{-1}(\psi) R I_{-1}^{dp}(\psi) \right]. \quad (2.21)$$

$$g_4(\psi) = \frac{1}{I} \sum_{j=2}^p \sum_{l=2}^p \left[I_{-1}^{dp}(\psi) \text{Trace} \left[I_{\beta}^{\psi}(\psi) \frac{\partial I_{\beta}^{\psi}(\psi)}{\partial g_1(\psi)} \right] \frac{\partial I_{\beta}^{\psi}(\psi)}{\partial g_1(\psi)} \right], \quad (2.20)$$

de dimensions semblables. En outre, $g_4(\psi)$ et $g_5(\psi)$ peuvent aussi s'écrire

$$y_i = X_i \beta + Z_i v_i + \varepsilon_i, \quad \varepsilon_i \sim N_m(0, R_i), \quad Z_i = (I - \rho W)^{-1}, \quad (3.1)$$

$$v_i = \kappa v_{i-1} + \eta_i, \quad \eta_i \sim N_m^m(0, \sigma_i^2 I) \quad i = 1, 2, \dots, T \quad \text{et} \quad (3.2)$$

À la présente section, nous utilisons des modèles d'espace d'états, obtenus au moyen de filtres de Kalman, pour exploiter les données chronologiques, ainsi que le paramètre de régression commun et le paramètre d'autocorrélation commun en vue de renforcer les estimateurs par sondage directs en tout point dans le temps. Cette approche est particulièrement avantageuse dans le cas où les estimations par sondage antérieures sont plus fiables. Les modèles utilisés dans cette catégorie sont les suivants

Ici, les paramètres ont la signification habituelle décrite à la section précédente. La matrice de poids $W(m \times m)$ et les matrices d'expérience $X_i(m \times p)$ sont connues, $Z(m \times m)$ est une matrice de coefficients des effets aléatoires et p est un coefficient d'autocorrélation inconnu. R_i est une matrice diagonale de dimension m qui peut être exprimée sous la forme $R_i = \text{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \dots, \sigma_{im}^2)$, où les σ_{i1}^2 sont les variances d'échantillonnage connues correspondant au i^{e} petit domaine et au i^{e} point dans le temps. β est un vecteur inconnu d'effets fixes et $\psi = [\rho, \sigma_1^2, \sigma_2^2, \kappa]^{\top}$ est un vecteur de trois paramètres inconnus. Ces paramètres sont indépendants du temps t . Il convient de souligner que les effets aléatoires v_i peuvent varier conformément à (3.2) et que κ est un paramètre temporel autorégressif. Pour la stationnarité, $|\kappa| < 1$.

l'erreur quadratique moyenne (EQM) du MPLNB comme suit

$$\theta(\psi) = X\beta(\psi) + A(\psi)[\gamma - X\beta(\psi)] \quad (2.5)$$

$$E[\theta(\psi) - \theta(\theta(\psi)) - \theta(\gamma)] = g_1(\psi) + g_2(\psi), \quad (2.6)$$

$$g_1(\psi) = \Delta(\psi)R = R - RZ^{-1}(\psi)R, \quad (2.7)$$

$$g_2(\psi) = RZ^{-1}(\psi)X(XZ^{-1}(\psi)X)^{-1}XZ^{-1}(\psi)R, \quad (2.8)$$

$$\begin{aligned} \beta(\psi) &= [XZ^{-1}(\psi)X]^{-1}XZ^{-1}(\psi)\gamma, \\ \Delta(\psi) &= \sigma^2 A^{-1}(\psi)Z^{-1}(\psi)A(\psi) = (I - pW)^T(I - pW). \end{aligned}$$

Ici, β , Δ et A sont tous des fonctions de ψ et sont habituellement exprimés sous la forme $\beta(\psi)$, $\Delta(\psi)$ et $A(\psi)$, respectivement. Cependant, dans certains cas, pour abréger, le suffixe ψ est omis. Le premier terme, $g_1(\psi)$, dans l'expression de l'EQM, montre la variabilité de θ quand tous les autres paramètres sont connus et est d'ordre $O(1)$. Le deuxième terme, $g_2(\psi)$, dû à l'estimation des effets fixes β , est d'ordre $O(m^{-1})$ pour les grandes valeurs de m . En outre, avec $p = 0$, le modèle susmentionné se réduit au modèle de régression linéaire à effets mixtes standard, tandis que, pour $X\beta = \mu$, nous obtenons un schéma purement spatial avec un terme d'ordonnée à l'origine uniquement.

En pratique, le paramètre ψ est inconnu et estimé d'après les données. Nous obtenons l'estimateur du maximum de vraisemblance (EMV) du paramètre ψ en maximisant la fonction de log-vraisemblance de ψ suivante

$$l = \text{const} - \frac{1}{2} \log[|\Sigma(\psi)|] - \frac{1}{2} [\gamma - X\beta(\psi)]^T \Sigma^{-1}(\psi) [\gamma - X\beta(\psi)] \quad (2.9)$$

par rapport au paramètre ψ . Nous obtenons le meilleur prédicteur linéaire sans biais empirique (MPLNBE), $\hat{\theta}(\psi)$ et l'estimateur naïf de l'EQM d'après les équations (2.5) et (2.6), respectivement, en remplaçant le vecteur de paramètres ψ par son estimateur $\hat{\psi}$.

$$\hat{\theta}(\psi) = \sigma^2 A^{-1}(\psi)Z^{-1}(\psi)\gamma + RZ^{-1}(\psi)X\hat{\beta}(\psi), \quad (2.10)$$

$$\text{EQM}[\hat{\theta}(\psi)] = g_1(\psi) + g_2(\psi), \quad (2.11)$$

$$\text{et } A(\psi) = (I - pW)^T(I - pW).$$

Cette expression de la EQM du MPLNBE sous-estime gravement la EQM réelle, car nous n'avons pas tenu compte

Conditions de régularité 1

de la variabilité due à l'estimation des paramètres d'après les données. Nous obtenons une approximation de deuxième ordre de $\text{EQM}[\hat{\theta}(\psi)]$ dans le cas où ψ est l'estimateur du maximum de vraisemblance restreint (EMVR) de ψ , en supposant que la valeur de m est grande et en négligeant tous les termes d'ordre $o(m^{-1})$, sous les conditions de régularité qui suivent. Nous avons calculé l'approximation en nous inspirant des méthodes de Prasad et Rao (1990) et de Datta et Lahiri (2000) qui sont de nature heuristique.

a) Les éléments de X sont bornés uniformément de sorte que $X^T Z^{-1}(\psi)X = [O(m)]^{p \times p}$, où $\Sigma(\psi) = [\sigma^2 A^{-1}(\psi) + R]$;

b) m est fini;

$$c) A(\psi)X = [O(1)]^{m \times p}, \quad (\partial A(\psi)X)/(\partial \psi^d) = [O(1)]^{m \times p}, \quad (\partial_z^2 V(\psi))/(\partial \psi^d \partial \psi^e) = [O(1)]^{m \times m} \text{ pour } d, e = 1, 2;$$

$$d) \psi \text{ est l'estimateur de } \psi \text{ qui satisfait } \psi - \psi = O_p(m^{-1/2}), \psi(-\gamma) = \psi(\gamma), \psi(\gamma + xh) = \psi(\gamma) \forall h \in R^p \text{ et } \forall \gamma.$$

Ces conditions de régularité sont satisfaites dans ce cas. La forme normalisée spéciale de la matrice de poids W satisfait la condition (c) pour $|p| > 1$, car elle ne contient qu'un nombre fini d'éléments non nuls et la somme des lignes est égale à 1. Il convient de mentionner ici que la matrice $\sigma^2 A^{-1} Z^{-1}$ contient un nombre fini d'éléments non nuls et que l'ordre de W , $(I - pW)$, $W(I - pW)Z^{-1}$ ou $W(I - pW)$, $W(I - pW)Z^{-1}$ est de leur ordre mentionnés à la condition (c) et de leurs déterminants mentionnés à la condition (c) n'augmente pas. En outre l'EMV et l'EMVR satisfont la condition (d). Nous montrons au théorème A.1 qu'une approximation de deuxième ordre de l'EQM du MPLNBE est

$$\text{EQM}[\hat{\theta}(\psi)] = E[\hat{\theta}(\psi) - \theta(\hat{\theta}(\psi)) - \theta(\gamma)] \quad (2.12)$$

$$= g_1(\psi) + g_2(\psi) + o(m^{-1}). \quad (2.13)$$

Ici, le troisième terme $g_3(\psi)$ provient de l'estimation du vecteur de paramètres inconnu provenant des données d'échantillon et est de même ordre $O(m^{-1})$ que $g_2(\psi)$. En outre, $g_3(\psi)$ peut s'exprimer sous la forme

$$g_3(\psi) = L^T(\psi)[L^{-1}(\psi) \otimes \Sigma(\psi)]L(\psi), \quad (2.13)$$

$$\text{où}$$

$$L(\psi) = \text{CO}^{PS2}[L^p(\psi), L^{o2}(\psi)]^T, L^p(\psi) = \frac{\partial \Delta(\psi)}{\partial \psi^p}, p = 1, 2, L^{o2}(\psi) = E\left[\frac{\partial^2 \psi \partial \psi^T}{\partial^2 \psi^2}\right]$$

Dans le contexte de l'Inde, le principal problème que posent les données est l'absence de données de registre administratives ou de l'état civil au niveau du petit domaine. Souvent, il est difficile de trouver des variables exogènes étroitement liées à la variable étudiée (coefficient de

corrélation multiple $R^2 > 0,5$).

Dans le présent article, nous envisageons l'exploitation d'une autocorrélation spatiale entre les petits domaines sous la forme d'un modèle spatial afin d'améliorer les estimateurs pour petits domaines. En outre, pour les données chronologiques, nous utilisons un modèle spatio-temporel du genre filtres de Kalman pour améliorer encore davantage les estimateurs. Nous étudions les données chronologiques sur les dépenses de consommation mensuelles par habitant (DCMH) estimées d'après une grande enquête par sondage réalisée par la National Sample Survey Organisation (NSSO). Dans le présent article, nous proposons des modèles appropriés dans le cadre du modèle linéaire à effets mixtes afin d'obtenir de meilleurs estimateurs des DCMH au niveau du petit domaine.

La présentation de la suite de l'article est la suivante. À la section 2, nous considérons un modèle spatial du genre du modèle linéaire à effets mixtes général avec introduction d'une autocorrélation spatiale entre les petits domaines. Nous présentons le meilleur prédicteur linéaire sans biais (MPLNB) et le meilleur prédicteur linéaire sans biais empirique (MPLNBE) des effets mixtes. Nous obtenons également une approximation de deuxième ordre de l'erreur quadratique moyenne (EQM) du MPLNBE et de l'estimateur de l'EQM. À la section 3, nous décrivons l'extension du modèle spatial aux séries chronologiques sous la forme d'un modèle spatio-temporel, suivant l'approche des filtres de Kalman. Nous discutons du MPLNB et du MPLNBE des effets mixtes, ainsi que d'une approximation de deuxième ordre de l'EQM du MPLNBE et de l'estimateur de l'EQM. À la section 4, nous présentons et analysons les estimations des DCMH provenant d'une grande enquête par sondage réalisée périodiquement en Inde. Enfin, à la section 5, nous présentons les conclusions de l'analyse des données. Toutes les preuves sont données en annexe.

2. Modèle spatial

Habituellement, les caractéristiques des petits domaines présentent une dépendance spatiale en ce qui concerne les similarités de quartier. Cressie (1990) a utilisé la dépendance spatiale conditionnelle pour les effets aléatoires dans le contexte de l'ajustement du sous-dénombrement au recensement. Ici, nous utilisons une dépendance spatiale simultanée (Cliff et Ord 1981) pour les effets aléatoires, qui présente certains avantages par rapport à la dépendance

1994).
 (2.1) $y = \theta + \varepsilon, \quad \varepsilon \sim N^m(0, R),$
 (2.2) $\theta = X\beta + u,$
 (2.3) $u = \rho W u + v, \quad v \sim N^m(0, \sigma_v^2 I),$
 où θ est un vecteur de dimension m (correspondant au nombre de petits domaines) pour la caractéristique étudiée et y est son estimateur par sondage direct obtenu au moyen de données recueillies auprès d'un petit échantillon. Dans le modèle susmentionné, la première équation (2.1) montre le modèle fondé sur le plan de sondage (échantillonnage), la deuxième (2.2) montre le modèle de régression et la troisième (2.3), le modèle spatial sur les résidus, les deux dernières étant liées dans la première équation. Le modèle susmentionné peut s'exprimer sous la forme

$$y = X\beta + Zv + \varepsilon, \quad Z = (I - \rho W)^{-1}, \quad (2.4)$$

où $X(m \times p)$ est la matrice d'expérience de plein rang colonne p , $\beta(p \times 1)$ est un vecteur-colonne de paramètres de régression et $Z(m \times m)$ représente les coefficients des effets aléatoires v . $W(m \times m)$ est une matrice de poids spatiaux connue qui montre le degré d'interaction dans toute paire de petits domaines. Les éléments de $W \equiv [W_{ij}]$ avec $W_{ii} = 0$ $\forall i$ peuvent dépendre de la distance entre les centres des petits domaines ou de la longueur de leur frontière commune. À titre de solution de rechange simple, elle peut avoir des valeurs binaires $W_{ij} = 1$ (non échelonnées) si le j^{e} domaine est physiquement contigu au i^{e} domaine et $W_{ij} = 0$, autrement. Nous avons normalisé la matrice de façon qu'elle satisfasse $\sum_{j=1}^m W_{ij} = 1$ pour $i = 1, 2, \dots, m$. La constante ρ est une mesure du niveau global d'autocorrélation spatiale et sa grandeur reflète la mesure dans laquelle W est appropriée sachant y et X . En outre, nous supposons que v et ε sont indépendantes l'une de l'autre. R est une matrice diagonale d'ordre m que l'on peut exprimer sous la forme $R = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2)$ où les σ_i^2 sont les variances d'échantillonnage connues correspondant au i^{e} domaine. Le vecteur de paramètres $\psi = [\rho, \sigma_v^2]^T$ contient deux éléments.

Ce modèle est renforcé par l'emprunt de données à des petits domaines similaires grâce à deux paramètres communs, c'est-à-dire le paramètre de régression β et le paramètre d'autocorrélation ρ . Notons que le présent modèle est plus général et qu'il permet d'obtenir le meilleur (Henderson 1975), nous pouvons obtenir le meilleur

Modèles spatio-temporels pour l'estimation pour petits domaines

Bharat Bhushan Singh, Girja Kant Shukla et Debasis Kundu¹

Résumé

Nous proposons un modèle de régression spatial dans un cadre général de modèles à effets mixtes pour résoudre le problème de l'estimation pour petits domaines. L'utilisation d'un paramètre d'autocorrélation commun à l'ensemble de petits domaines permet de produire de meilleures estimations pour petits domaines. Ce paramètre s'avère fort utile dans les cas où l'utilisation de variables exogènes améliore peu ces estimations. Nous élaborons également une approximation de deuxième ordre de l'erreur quadratique moyenne (EQM) du meilleur prédicteur linéaire sans biais empirique (MPLNBE). En suivant l'approche des filtres de Kalman, nous proposons un modèle spatio-temporel. Dans ce cas également, nous obtenons une approximation de deuxième ordre de la BQM du MPLNBE. À titre d'étude de cas, nous utilisons les données de la série chronologique sur les dépenses de consommation mensuelles par habitant (DCMH) provenant de la National Sample Survey Organisation (NSSO) du ministère de la Statistique et de la Mise en œuvre des programmes du gouvernement de l'Inde pour valider les modèles.

Mots clés : Modèle linéaire à effets mixtes; autocorrélation spatiale; matrice de poids; meilleur prédicteur linéaire sans biais; meilleur prédicteur linéaire sans biais empirique; filtres de Kalman; cycles de la NSSO.

1. Introduction

La planification au niveau local nécessite des données fiables de niveau approprié. La réalisation de recensements complets ou de grandes enquêtes par sondage auprès d'un échantillon de taille adéquate est coûteuse et longue. Les recensements sont généralement réalisés une fois tous les dix ans, tandis que les enquêtes par sondage sont souvent planifiées pour fournir des estimations à un niveau beaucoup plus élevé d'aggrégation. L'une de ces grandes enquêtes par sondage est l'enquête socioéconomique de la National Sample Survey Organisation (NSSO). Ici, les estimations par sondage directes sont disponibles au niveau du petit domaine (district), car la plupart des districts reprennent une strate dans la procédure d'échantillonnage adoptée par la NSSO. Cependant, ces estimations sont très peu fiables, à cause d'erreurs-types inacceptablement grandes. Il est donc nécessaire de les renforcer à l'aide d'information provenant de petits domaines semblables ou de variables exogènes avec lesquelles un lien peut être établi, faciles à obtenir et reliées à la variable étudiée.

Diverses approches fondées sur un modèle ont été proposées pour améliorer les estimateurs directs. L'approche fondée sur un modèle facilite la validation au moyen de données d'échantillon. Le modèle simple caractéristique du domaine qui est proposé est le modèle à deux degrés de Fay et Herriot (1979).

$$y_i = \theta_i + \varepsilon_i, \quad E(\varepsilon_i | \theta_i) = 0, \quad \text{Var}(\varepsilon_i | \theta_i) = \sigma_{\varepsilon}^2, \quad (1.1)$$
$$\theta_i = X_i' \beta + v_i z_i, \quad E(v_i) = 0, \quad \text{Var}(v_i) = \sigma_v^2, \quad i=1, 2, \dots, m, \quad (1.2)$$

Ici, les y_i sont des estimateurs directs par sondage des θ_i , des caractéristiques étudiées. Les θ_i peuvent être les moyennes de petit domaine dans la population. Les $X_i' = (X_i^{(1)}, \dots, X_i^{(p)})^T$ sont des variables exogènes qui sont disponibles et que l'on suppose être étroitement liées aux θ_i et les z_i sont des constantes positives connues. β ($p \times 1$) est le vecteur des paramètres de régression.

La première équation (1.1) est le modèle fondé sur le plan de sondage, tandis que la deuxième (1.2) est le modèle de lien. Les ε_i sont les erreurs d'échantillonnage et les v_i sont des variables aléatoires indépendantes et de même loi (iid). On suppose souvent que les erreurs et les effets aléatoires suivent une loi normale. Pour ce modèle, nous proposons le meilleur prédicteur linéaire sans biais (MPLNB) selon le modèle du meilleur estimateur linéaire sans biais (MELB). L'estimation est convergente par rapport au plan et sans biais par rapport au modèle (Chosh et Rao 1994). Il s'agit typiquement de la moyenne pondérée de l'estimateur par sondage direct y_i et de l'estimateur synthétique par la régression $X_i' \beta + v_i z_i$. L'estimateur MPLNB dépend de la composante de la variance σ_{ε}^2 qui est inconnue dans les applications pratiques. Diverses méthodes d'estimation des composantes de la variance σ_{ε}^2 dans le modèle linéaire à effets mixtes général existent (Cressie 1992). En remplaçant σ_{ε}^2 par un estimateur asymptotique-moment convergent $\hat{\sigma}_{\varepsilon}^2$, nous obtenons également un meilleur prédicteur linéaire sans biais empirique (MPLNBE).

- O'Malley et Zaslavsky : Fonctions de variance-covariance pour les moyennes de domaine
- Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.
- Spencer, B.D. (2000). Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection. *Techniques d'enquête*, 26, 153-155.
- Valliant, R. (1992a). Longitudinal smoothing of price index variances. Dans *Statistics Canada Symposium*. Ottawa: Statistique Canada, 113-120.
- Valliant, R. (1992b). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 8, 433-444.
- Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Woodruff, S. (1992). Variance estimation for estimates of employment change in the Current Employment Statistics Survey. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA: American Statistical Association, 626-631.
- Zaslavsky, A.M., Beaulieu, N.D., Landon, B.E. et Cleary, P.D. (2000). Dimensions of consumer-assessed quality of Medicare managed-care health plans. *Medical Care*, 38, 162-174.
- Zaslavsky, A.M., et Cleary, P.D. (2002). Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 Survey. *Medical Care*, 40, 951-964.
- Ståndal, C.-E., Swensson, B. et Wretling, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Goldstein, E., Cleary, P.D., Langwell, K.M. Zaslavsky, A.M. et Heller, A. (2001). Medicare Managed Care CAHPS: A tool for performance improvement. *Health Care Financing Review*, 22, 101-107.
- Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.
- Hays, R.D., Shaul, J.A., Williams, V.S.L., Lubalin, J.S., Harris-Kojetan, L.D., Sweeny, S.F. et Cleary, P.D. (1999). Psychometric properties of the CAHPS 1.0 survey measures. *Medical Care*, 37 (Supplement), 22-31.
- Huff, L.L., Ellinger, J.L. et Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. Current Employment Survey. Dans *Proceedings of the Joint Statistical Meetings* [CDROM], Alexandria, VA: American Statistical Association, 1519-1524.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.
- O'Malley, A.J. et Zaslavsky, A.M. (2004). Implementation of cluster-level covariance analysis for survey data with structured nonresponse. Dans *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 1907-1914.
- Otto, M.C., et Bell, W.R. (1995). Sampling error modeling of poverty and income statistics for states. Dans *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.
- Ståndal, C.-E., Swensson, B. et Wretling, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

type 1-4, les corrélations pourraient tout aussi bien être modélisées par des constantes, ce qui permet aussi de garantir plus facilement le caractère défini positif de la matrice de corrélations prédite. Cependant, il est important que les paramètres du modèle de corrélation puissent varier selon la paire de questions.

Un estimateur composite, résultant de la combinaison pondérée des estimateurs direct et fondé sur un modèle proportionnellement à leur précision, a une variance plus faible que l'un et l'autre estimateur pris individuellement, surtout quand les composantes ont des poids presque égaux. L'estimateur fondé sur un modèle est celui dont l'influence sur les estimations pour les petits domaines pour lesquels on dispose de peu d'information est la plus forte. L'influence de l'estimateur fondé sur un modèle est, par ordre décroissant d'importance, la plus forte sur les estimations de la variance, les corrélations des questions de même type et, enfin, les corrélations des questions de types différents. Les estimateurs fondés sur un modèle et les estimateurs composites peuvent les uns et les autres étre calés (ajustement par le quotient) de sorte que les moyennes sur l'ensemble des domaines concordent avec les estimations directes, bien que cela n'ait pas été nécessaire dans notre exemple.

Les fonctions de variance et de covariance généralisées (FVCG) ont plusieurs applications dans nos travaux de recherche en cours. Nous élaborons des méthodes fondées sur la quasi-vraisemblance d'estimation des matrices de covariance pour les moyennes de domaines des questions ordonnées d'enquête, avec représentation de la covariance de deuxième niveau (structurale) au moyen d'un modèle hiérarchique (O'Malley et Zaslavsky 2004). Les modèles de FVCG sont nécessaires pour obtenir des estimations des variances et des covariances d'échantillonnage, ainsi que pour modifier ces estimations lorsque les moyennes sont restituées durant la procédure d'ajustement de modèle. Si la variabilité d'échantillonnage des estimations des FVCG est minimale parce que le nombre de domaines est grand, les variances et les covariances prévues des FVCG peuvent être considérées comme étant connues. Cependant, si l'erreur d'échantillonnage des estimations fondées sur les FVCG est importante, il convient d'utiliser un modèle permettant à ces erreurs de se propager tout au long de l'analyse. Dans le cadre de travaux connexes, Fay et Train (1997) ont utilisé un modèle binomial avec un effet de plan pour chaque domaine dans l'estimation bayésienne empirique des taux binomiaux. Notre étude étend cette approche à l'estimation multivariée et à des formats de réponse plus généraux.

Une autre application des FVCG est le calcul des estimations de la variance pour les combinaisons linéaires de moyennes de question afin de faciliter l'estimation de la variance de scores composites, comme ceux utilisés dans la

Bibliographie

Choi, M.J., Elling, J.L., Gershunskaya, J. et Huff, L.T. (2002). Evaluation of generalized variance function estimators for the U.S. Current Employment Survey. Dans *Proceedings of the Joint Statistical Meetings* [CDROM]. Alexandria, VA: American Statistical Association, 534-539.

Elling, J. (2002). Use of generalized variance functions in multivariate analysis. Dans *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, 904-913.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fay, R.E., et Train, G.F. (1997). Small domain methodology for estimating income and poverty characteristics for states in 1993. Dans *Proceedings of the Social Statistics Section*. Alexandria, VA: American Statistical Association, 183-188.

Freund, J.E., et Walpole, R.E. (1987). *Mathematical Statistics*. New Jersey: Prentice-Hall, Inc., 4^{ème} Edition.

Gabler, S., Haeder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Statistique Canada, N° 12-001-XPB au catalogue

Remerciements

Les présents travaux ont été financés par la U.S. Agency for Healthcare Research and Quality par la voie de la Consumer Assessments of Health Plans Study (subvention U18 HS09205-06) et par les U.S. Centers for Medicare and Medicaid Services (contrat 500-95-007). Nous remercions Paul D. Cleary de son appui continu durant les travaux, Matt Cioffi de la gestion des données, ainsi qu'Elizabeth Goldstein et Amy Heller des Centers for Medicare and Medicaid Services (CMS), et les autres membres de l'équipe de mise en œuvre de l'Enquête CAHPS-MMC.

La méthodologie des FVCG peut étre étendue de plusieurs façons. Outre les mesures sommatores des résultats, les fonctions de variance et covariance généralisées (FVCG) peuvent dépendre d'autres variables indépendantes, en particulier celles qui prédiraient mieux les corrélations. Nous avons considéré des variables résumant les profils de réponse, comme la proportion de répondants dans un domaine, mais celles-ci n'ont pas amélioré le modèle. Les FVCG pourraient aussi étre étendues à l'échantillonnage à plusieurs degrés.

4.6 Prédiction conjointe

questions de type 0-10 ou 1-4, parce que les matrices de corrélations prédites sont définies positives pour chaque domaine.

5. Conclusion

Nous présentons la méthodologie pour estimer les fonctions de variance et de covariance pour les moyennes de domaine de questions d'enquête ordonnées. Notre méthodologie peut également s'appliquer aux questions d'enquête comportant une échelle de mesure continue. Nous présentons une décomposition de l'erreur de modélisation qui permet de séparer la variation due à l'échantillonnage de celle due à l'ajustement du modèle. La décomposition permet aussi d'éviter le surajustement du modèle, parce qu'elle estime la proportion de la variation des données qui peut être modélisée et, donc, le moment où les prédicteurs courants suffisent.

La procédure d'ajustement des modèles de variance et de corrélation est la même que les données contiennent ou non des enchevêtrements structuraux de questions. L'exposé analytique de la section 3.3 montre que, s'il existe des enchevêtrements de questions, il faut connaître les différences moyennes entre les questions selon la situation de réponse à d'autres questions afin d'estimer la covariance d'échantillonnage. Cependant, nous soutenons que ces quantités ont, vraisemblablement un effet minime sur les résultats et que, par conséquent, on pourrait utiliser un modèle constant, argument qui est appuyé par nos résultats empiriques.

Une fonction de variance quadratique dont la valeur est contrainte d'être nulle pour l'évaluation maximale et un modèle pour les corrélations transformées comportant le produit, mais non le carré des moyennes sont les meilleurs prédicteurs des estimations directes dans l'application que nous avons choisie comme exemple. En général, les erreurs-types des estimations modélisées de la variance sont beaucoup plus faibles que celles des estimations directes; toutefois, il n'en est pas ainsi des estimations des corrélations. Il est intéressant et rassurant de constater que notre fonction de variance quadratique peut être exprimée sous la forme du modèle de variance relative très répandu de Wolter (1985).

Pour nos données ordonnées, les évaluations moyennes de domaine contiennent des informations minimales concernant la corrélation entre les évaluations. Donc, la relation moyenne-covariance est principalement un artefact de la relation moyenne-variance. Cependant, pour les questions comportant un grand nombre de catégories de réponse, l'association entre les corrélations et les résultats moyens pour les questions de même type est plus forte, surtout pour les paires de questions à échelle 0-10. À part les évaluations de type 0-10 et, éventuellement, celles de

Comme nous avons modélisé indépendamment les corrélations pour chaque question, nos matrices de corrélations ajustées ne satisfont pas nécessairement la contrainte de

définie positive, qui peut être importante pour les inférences multivariées. Dans le cadre de travaux supplémentaires, nous avons déterminé qu'à condition de limiter l'analyse multivariée aux questions de même type, les corrélations ajustées d'après les modèles C2 et C4 donnent des estimations définies positives des matrices de corrélation pour presque tous les domaines. Cependant, pour les analyses portant sur des questions de types différents (par exemple, les questions à échelle numérique 0-10 et les questions à réponse oui/non 1-2), les prédictions fondées sur C4 donnent des matrices de corrélations qui sont indéfinies pour de nombreux domaines, tandis que celles fondées sur C2 sont plus stables et donnent presque systématiquement des matrices définies positives. Ceci donne à penser que, si

C4 peut être légèrement supérieur en ce qui concerne l'ajustement du modèle univarié, C2 pourrait être plus approprié pour l'inférence multivariée.

Un moyen de contourner le problème des matrices de corrélations prédites indéfinies consiste à utiliser une moyenne pondérée de la matrice de corrélations prédite pour un domaine et de la matrice de corrélations moyenne estimée (MCMBE) sur l'ensemble de domaines. Nous pouvons construire la MCMBE par pondération des estimations directes (chacune étant au moins semi-définie positive) par la taille totale de l'échantillon pour chaque domaine. Puis, nous remplaçons toute matrice de corrélations prédite indéfinie par la moyenne pondérée de la matrice de corrélations prédite et de la MCMBE, en accordissant le poids utilisé pour chaque domaine jusqu'à ce que nous obtenions une matrice définie positive. Comme pour un estimateur bayésien empirique, ce processus stabilise les estimations en réduisant effectivement les coefficients du modèle vers ceux d'un modèle plus simple (constant).

Lors de l'analyse simultanée des 35 questions de la CAHPS, la MCMBE avait un poids moyen sur l'ensemble des domaines de 0,65 pour le modèle C4, mais de 0,01 seulement pour le modèle C2, puisque les corrélations prédites sous C2 sont habituellement définies positives. Lors de l'analyse des questions de type 0-10, 1-4 ou 1-3 seulement, la MCMBE avait un poids moyen de 0,28 et de 0,00 pour C4 et C2, respectivement, tandis que lors de l'analyse des questions de type 0-10 ou 1-4 seulement, les poids moyens correspondants étaient de 0,06 et de 0,00. Lors de l'analyse de questions types de questions séparément, le poids moyen de la MCMBE avec le modèle C4 était de 0,00 pour les questions de type 0-10 ou 1-4, de 0,01 pour les questions de type 1-3 et 0,17 pour les questions de type 1-2. La MCMBE n'est donc pas nécessaire pour analyser les

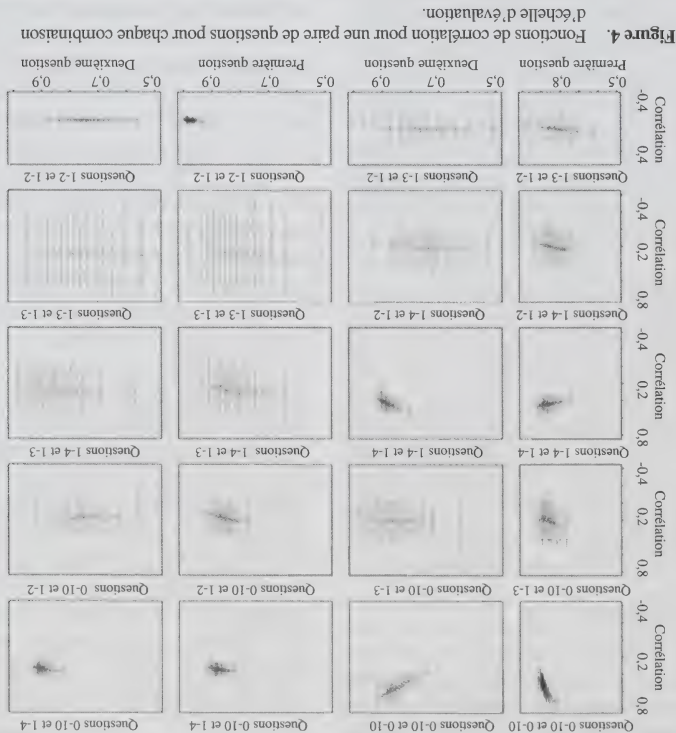


Tableau 7
Distribution des poids pour la composante fondée sur un modèle de l'estimateur composite, moyenne sur les questions de même type

Modèle	Type de question		Prob(Modèle < Variation	Quantiles
	1	2		

Variance	Corrélation	d'échantillonnage	
		10 %	Médiane
0-10	1-4	0,983	0,974
1-3	1-4	0,996	1,000
0-10	0-10	0,038	0,256
1-4	0-10	0,358	0,468
1-3	0-10	0,523	0,540
1-2	0-10	0,784	0,695
1-4	1-4	0,435	0,497
1-3	1-4	0,516	0,587
1-3	1-3	0,827	0,737
1-2	1-3	0,799	0,709
1-2	1-2	0,541	0,584
1-2	1-2	0,799	0,709

La distribution des poids est résumée au moyen des 10^e, 50^e et 90^e centiles. Voir le tableau 3 pour la définition de ModEtr.

colinéarité entre les prédicteurs dans le modèle C4. Dans de nombreux cas, la valeur estimée de α_4 contrebalance les estimations des paramètres des prédicteurs linéaires, ce qui donne une courbe ajustée pratiquement plate.

4.4 Fonctions de différence de moyennes

La différence $D_{h,ij}$ semble ne dépendre ni de la moyenne marginale ni de son carré, ce qui implique qu'un modèle analogue à V3 pourrait être approprié. Cependant, comme $D_{h,ij}$ est habituellement suffisamment faible pour que $D_{h,ij} D_{h,ij}$ ait un effet minime sur (16), nous ajustons un modèle constant.

4.5 Estimateur composite

Le tableau 7 donne les valeurs moyennes, calculées sur l'ensemble des questions (ou paires de questions) de même type, des quantiles de la distribution des poids $\sigma_h^2/(\tau^2 + \sigma_h^2)$ pour l'estimation fondée sur un modèle utilisée dans l'estimateur composite de la section 3.5. La proportion fondée sur un modèle est plus faible que celle des estimations directes est également présentée. Comme nous l'avons mentionné plus haut, les prédictions fondées sur un

Tableau 5

Diagnostiques d'ajustement de modèle pour les fonctions de corrélation pour les questions de même type, moyenne sur les paires de questions de même type

Échelle d'évaluation		0 - 10		1-4		1-3		1-2	
Variation d'échantillonnage		ModEtr	R ²	ModEtr	R ²	ModEtr	R ²	ModEtr	R ²
Modèle C1	Modèle C1	0,060	0,000	0,028	0,000	0,112	0,000	0,018	0,000
Modèle C2	Modèle C2	0,060	0,013	0,025	0,070	0,103	0,048	0,017	0,014
Modèle C3	Modèle C3	0,057	0,039	0,024	0,079	0,102	0,054	0,017	0,018
Modèle C4	Modèle C4	0,047	0,150	0,023	0,100	0,100	0,068	0,016	0,029
Modèle C5	Modèle C5	0,044	0,151	0,105	0,096	0,096	0,080	0,015	0,034
Prob(ModEtr < Variation d'échantillonnage)		0,339		0,399		0,461		0,788	
Modèle C1	Modèle C1	0,033	0,033	0,400	0,411	0,498	0,795	0,796	0,96
Modèle C2	Modèle C2	0,033	0,034	0,400	0,411	0,498	0,795	0,796	0,96
Modèle C3	Modèle C3	0,038	0,038	0,435	0,516	0,516	0,799	0,799	0,802
Modèle C4	Modèle C4	0,038	0,038	0,440	0,530	0,530	0,802	0,802	0,802
Prob(ModEtr < Variation d'échantillonnage)		0,358		0,435		0,523		0,784	
Modèle C5	Modèle C5	0,065	0,065	0,440	0,530	0,530	0,802	0,802	0,802
Voir le tableau 1 pour une description des questions de type 0-10, 1-4, 1-3 et 1-2, et le tableau 3 pour une explication des en-têtes de colonne.									

Tableau 6

Diagnostiques d'ajustement de modèle pour les fonctions de corrélation pour C4 selon le type de question. Moyenne sur les questions de même type.

Type	0-10		1-4		1-3		1-2	
	ModEtr	R ²	ModEtr	R ²	ModEtr	R ²	ModEtr	R ²
-0-10	0,047	0,149	0,021	0,104	0,040	0,094	0,013	0,059
-1-4			0,023	0,100	0,038	0,076	0,013	0,039
-1-3					0,100	0,068	0,028	0,031
-1-2							0,016	0,029
Prob(ModEtr < Variation d'échantillonnage)	0,358		0,435		0,523		0,784	
-0-10	0,038				0,605		0,790	
-1-4					0,516		0,827	
-1-3							0,799	
-2								

Dans la plupart des cas, les coefficients pour les termes $d^{h,i}$ ainsi que ceux de $d^{h,i}(1-d)^{h,i}$ de V_3 sont significatifs, ce qui indique que ces termes sont nécessaires pour la modélisation généralisée de la variance. Dans certains cas (particulièrement pour les questions à échelle 0-10), le coefficient du terme $d^{h,i}(1-d)^{h,i}$ est négatif, ce qui donne une fonction de variance estimée convexe plutôt que concave (forme de la fonction de variance binomiale). Cette situation peut se produire si les moyennes d'échantillon de ces évaluations sont concentrées sur une petite partie de l'échelle de réponses, sur laquelle le terme linéaire explique une grande part de la variation des données. Comme nous l'avons mentionné plus haut, l'ajout de fonctions polynomiales ou logarithmiques d'ordre plus élevé de $d^{h,i}$ n'améliore pas significativement l'ajustement du modèle.

Tableau 4
Estimations moyennes des paramètres de la fonction de variance pour chaque type de question et écart-type entre les questions (entre parenthèses)

Modèle	0-10			1-4			1-3		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
V1	0,236 (0,016)	—	0,334 (0,039)	—	—	0,566 (0,068)	—	—	—
V2	—	0,271 (0,020)	—	—	0,421 (0,034)	—	—	0,711 (0,069)	0,420 (0,110)
V3	0,334 (0,143)	—0,114 (0,155)	0,151 (0,104)	—	0,241 (0,132)	—	—	0,239 (0,112)	—

et 1-3.

4.3 Fonctions de corrélation

Les modèles sont classés du plus simple (C1, modèle constant) au plus complexe (C5, contenant tous les termes linéaires et quadratiques). Comme pour les modèles de variance, les tests statistiques indiquent des effets d'interaction entre questions hautement significatifs, qui sous-entendent que des modèles distincts devraient être ajustés pour chaque paire. Nous ne nous attendions pas à ce que toutes les paires de questions présentent les mêmes corrélations, puisque, intentionnellement, les questions sont réparties en groupes intentionnellement cohérents, qui mesurent chacun un aspect distinct des expériences vécues par les patients, comme les interactions avec le médecin ou celles avec les agents des services à la clientèle (Hays, Shaul, Williams, Lubalin, Harris-Kojetan, Sweeney et Cleary 1999). Les ajustements des modèles de corrélation pour les

paire de questions de même type sont résumées au tableau 3. Pour la gamme de modèles considérée, les améliorations les plus importantes de l'efficacité des modèles (mesurées par R^2) sont celles observées entre C1 et C2, et entre C3 et C4. Par exemple, le R^2 moyen pour les évaluations numériques à échelle 1-10 dans les modèles C3 à C5 est de 0,0391, 0,1494 et 0,1508, respectivement, et le R^2 moyen pour les

évaluations à échelle 1-4 pour les modèles C1 à C3 est de 0,0700 et 0,0789, respectivement. Ces résultats donnent à penser que C2 et C4 sont les meilleurs modèles pour différencier paires de questions, affirmation qui est appuyée par les tests de vérification d'hypothèse concernant la signification des améliorations marginales de l'ajustement des modèles. La variation d'échantillonnage la plus élevée s'observe pour les échelles d'évaluation 1-3, du moins en partie parce que les taux élevés de non-réponses dues à des enchevêtrements de questions ont réduit les tailles d'échantillon. L'erreur de modélisation et le R^2 des modèles de corrélation pour les questions de types différents sont semblables à ceux des modèles pour les questions de même type. Les valeurs de R^2 des modèles de corrélation sont comprises entre 0,029 et 0,15 pour toutes les paires de questions. Bien qu'il n'existe aucune preuve que C4 soit un modèle inapproprié pour les corrélations, ces résultats indiquent qu'une variation importante des corrélations ne peut être

Les variances d'échantillonnage des estimations directes sont souvent inférieures aux variances de l'erreur de modélisation correspondante (partie inférieure des tableaux 5 et 6, particulièrement pour les questions à échelle 0-10, Sous C4, les variances de l'erreur de modélisation ne sont plus faibles que pour 13 % des domaines pour les évaluations de type 0-10, 40 % des domaines pour les évaluations de type 1-4 et environ 18 % des domaines pour les évaluations de type 1-3 ou 1-2.

La figure 4 donne les corrélations observées et la fonction ajustée C4 pour un exemple de paire de questions pour chacune des dix combinaisons de type de question, représentant les 55 paires de questions distinctes. Pour illustrer les modèles de corrélation ajustés, nous rajustons les corrélations observées et ajustées sur la moyenne de l'une des questions et représentons graphiquement les valeurs résultantes dans un espace bidimensionnel. Nous réplétons le processus pour l'autre question, ce qui nous donne deux traces pour chaque corrélation.

La figure 4 illustre la relation généralement faible entre les corrélations et les moyennes des questions observées aux tableaux 5 et 6. L'analyse de ces deux tableaux révèle que la relation entre la corrélation et le résultat moyen est plus faible pour les questions comportant un petit nombre de catégories et pour les corrélations de questions de différents types. En particulier, les évaluations numériques à échelle 0-10 sont le seul groupe pour lequel il existe une relation claire corrélation-moyenne.

Bien que les courbes ajustées pour les fonctions de

corrélation soient presque plates, la variation des estimations des paramètres sous le modèle C4 pour α_4 sont grandes et évoquent une instabilité. La très forte variabilité des estimations des paramètres est une conséquence de la

La partie inférieure du tableau 3 donne, pour chaque question, la proportion de domaines (parmi ceux comptant au moins deux réponses à la question considérée) pour lesquels la variation d'échantillonnage est supérieure à la variation de l'erreur de modélisation. Pour plus de 90 % des domaines, la variation de l'erreur de modélisation est plus faible que la variation d'échantillonnage de l'estimation directe de la variance.

La figure 3 illustre l'ajustement de V3 pour deux questions de chacun des groupes 0-10, 1-4 et 1-3. Les illustrations pour les autres questions sont semblables, mais ne sont pas présentées ici faute d'espace. Les courbes ajustées sont contrainues de passer par la valeur 0 à l'évaluation maximale. Pour évaluer l'effet de cette contrainte sur la fonction de variance ajustée, nous ajustons également une fonction de variance quadratique (à trois paramètres) non contrainte. Celle-ci atteint une valeur très proche de 0 à l'évaluation maximale et s'approche de très près de la courbe ajustée d'après les modèles avec contraintes, ce qui appuie encore davantage le modèle V3.

Les estimations moyennes des paramètres et de leurs écarts-types sur l'ensemble des questions de même type sont présentés au tableau 4. La valeur des paramètres varie considérablement selon la question, ce qui soutient la décision d'estimer des coefficients de régression distincts.

R^2 pour le modèle V3 s'approchant de 0,75 pour les questions à échelle numérique (0-10), de 0,85 pour les questions à échelle de fréquence (1-4) et de 0,95 pour les questions à échelle d'intensité de problème (1-3).

Tableau 3

Statistiques de qualité d'ajustement du modèle pour les fonctions de variance									
Échelle d'évaluation	0-10	1-4	1-3	d'échantillonnage					
				ModEtt	R^2	ModEtr	R^2	ModEtr	R^2
Modèle V1	0,020	0,741	0,066	0,824	0,069	0,916			
Modèle V2	0,043	0,710	0,036	0,835	0,000	0,940			
Modèle V3	0,016	0,750	0,024	0,847	0,000	0,947			
d'échantillonnage									
Modèle V1	0,968	0,916	0,996						
Modèle V2	0,858	0,967	0,996						
Modèle V3	0,981	0,983	0,996						

ModEtr est la composante de la variance due au manque d'ajustement, R^2 est la statistique définie à la section 3.4, Prob(ModEtr < Variation d'échantillonnage) est la proportion de domaines pour lesquels l'erreur de modélisation est plus faible que la variation d'échantillonnage. Toutes les évaluations sont rééchantillonnées sur une échelle de 0-1 et les erreurs de modélisation sont multipliées par 10⁴.

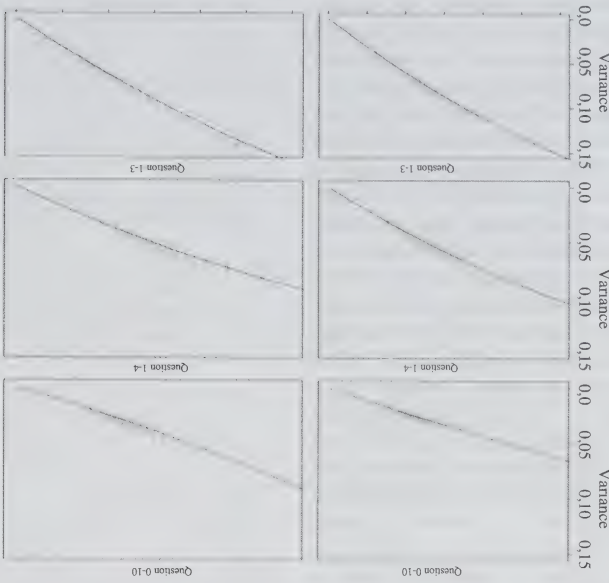


Figure 3.

Fonction de variance quadratique (V3) de deux questions pour chaque type d'évaluation. Chaque point est la moyenne de 60 domaines. Les lignes verticales joignent les 10^e et 90^e centiles de la distribution des variances. Pour ces tracés et les suivants, la direction de l'axe horizontal a été inversée afin qu'elle concorde avec celle des variables originales.

Nous présentons les modèles ajustés aux variances et aux corrélations dans la suite de la section. Des vérifications approfondies des modèles les mieux ajustés ont montré que les résidus ne présentaient aucune régularité discernable.

4.2 Fonctions de variance

Lors de travaux préliminaires non présentés ici, nous avons ajusté deux modèles pour les groupes de questions ayant la même échelle de réponse. Un avec les mêmes paramètres de régression pour toutes les questions et l'autre avec des paramètres de régression différents pour chaque question, à l'ensemble de données comprenant toutes les questions. Les comparaisons des ajustements globaux des modèles (au moyen de critères tels que le C_p de Mallows, le R^2 et le R^2 corrigé) et les tests de signification des interactions effet-question ont montré que permettre aux paramètres de varier selon la question améliore de façon significative l'ajustement du modèle. Par exemple, pour les évaluations numériques réchelonées, pondérées par la taille d'échantillon du domaine, les racines des erreurs quadratiques moyennes des deux modèles étaient de 0,446 contre 0,402, et les valeurs de R^2 étaient de 0,783 contre 0,825. D'après ces résultats, nous avons décidé d'ajuster des modèles distincts pour chaque question.

Nous avons ajusté les fonctions de variance (8) à (10) à chaque question, sauf celles à réponse oui/non, qui suivent la fonction de variance binomiale dans le cas de l'échantillonnage avec probabilités égales. La procédure itérative décrite à la section 3.4 a convergé presque précisément en exactement deux itérations. Ce résultat tient au fait que les poids des observations ne varient qu'en fonction de l'estimation de τ^2 , de sorte que leur variation est très faible après la première itération.

Le tableau 3 donne la variation d'échantillonnage moyenne, la variation moyenne de l'erreur de modélisation (ModelErr) et R^2 pour la valeur moyenne de chaque modèle calculés sur l'ensemble des questions correspondant à chaque type d'échelle. La variation d'échantillonnage, calculée selon (19), ne dépend pas du modèle.

Pour les questions ne comportant qu'un petit nombre de catégories (celles qui ressemblent le plus à la loi binomiale), la composante quadratique de la fonction de variance a tendance à dominer la composante linéaire, ce qui produit un meilleur ajustement des modèles V2 et V3 que du modèle V1. Comme V2 impose une contrainte en un point très en dehors de la fourchette de valeurs des moyennes de domaine, il n'est pas aussi bien ajusté aux données quand le nombre de catégories augmente et que les données s'écartent par conséquent davantage de la loi binomiale. Les réponses aux questions à échelle 0-10 sont moins dispersées que celles aux questions à échelle 1-4 ou 1-3, de sorte que le modèle linéaire est mieux ajusté. Les valeurs de

Les corrélations d'échantillon varient également beaucoup d'une paire de questions à l'autre (figure 2), quoique la plupart soient positives. Le plus souvent, les corrélations entre questions de même type sont plus fortes que celles entre questions de types différents. Les évaluations numériques à échelle 0-10 sont celles dont les corrélations sont les plus importantes (moyenne = 0,49) et, en général, les évaluations comportant un grand nombre de catégories ont tendance à produire des corrélations plus fortes que celles comportant un moins grand nombre de catégories. Bien que la plupart des paires de questions à échelle 1-4 donnent des corrélations moyennes s'approchant de 0,5, l'une des questions est négativement corrélée aux autres (révélée par la grappe de corrélations moyennes inférieures à 0); ce résultat est dû au codage inversé d'une question dont la moyenne globale d'échantillon ne se situait pas dans la moitié supérieure de l'échelle. Les distributions des corrélations des paires de questions dichotomiques 1-2 sont centrées autour de 0, ce qui signifie que ces corrélations sont souvent négatives. L'énoncé complet des questions et des statistiques sommaires supplémentaires figure dans Zaslavsky, Beaulieu, Landon et Cleary (2000) et dans Zaslavsky et Cleary (2002).

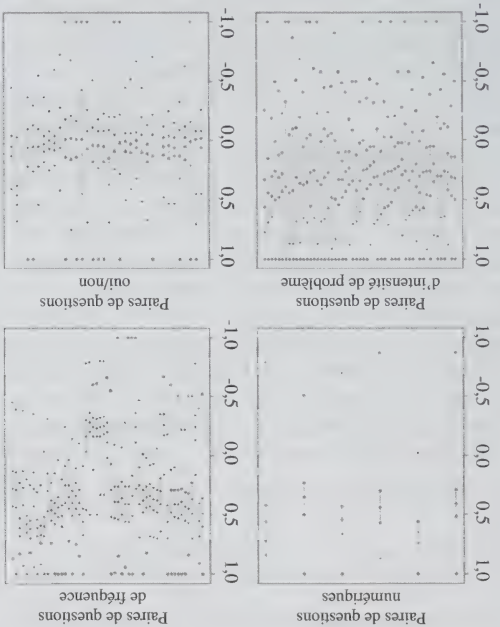


Figure 2. Sommaire de cinq points des corrélations d'échantillon de domaine entre les questions de même type. Le sommaire de cinq points comprend le minimum, le 10^e centile, la moyenne, 90^e centile et le maximum.

pour une échelle donnée. Les questions comportant un plus grand nombre de catégories de réponse sont concentrées vers l'extrémité supérieure de l'échelle et ont donc une variance plus faible. Par exemple, l'écart-type moyen pour les questions à échelle 1-2 (0,36) est égal au double de celui des questions à échelle 0-10, les distributions des évaluations moyennes de domaine varient fortement entre les questions de même type. Par exemple, l'écart-type des moyennes des questions à échelle 1-2 sur l'ensemble des questions est de 0,30 comparativement à un écart-type rééchantonné de 0,03 pour les questions à échelle 0-10.

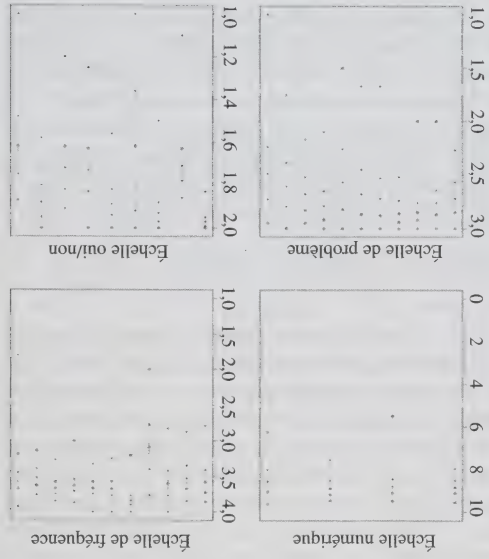


Figure 1. Sommaire de cinq points des moyennes d'échantillon de domaine pour chaque type de question. Le sommaire de cinq points comprend le minimum, le 10^e centile, la moyenne, le 90^e centile et le maximum.

Tableau 2

Statistiques sommaires des moyennes et des écarts-types de domaine évalués sur les domaines et sur les questions

Type	Statistiques sommaires des moyennes et des écarts-types de domaine évalués sur les domaines et sur les questions	E.T. des questions
Min	Moy.	Moy.
Numérique 0-10	6,82	9,52
Fréquence 1-4	2,86	3,90
Problème 1-3	1,88	2,99
Oui/non 1-2	1,34	1,96

Nota : Les colonnes 2 à 5 donnent le minimum, le maximum, la moyenne et l'écart-type des moyennes des questions de domaine sur l'ensemble des questions d'un type donné. Les colonnes 6 et 7 donnent la moyenne et l'écart-type des écarts-types des questions de domaine sur l'ensemble des questions d'un type donné.

réponses assez faibles. La variation la plus forte des proportions de questions saines s'observe pour les questions à réponse oui/non, le taux allant de 96,7 % pour « une plainte ou un problème concernant le régime » à 12,5 % pour « l'obtention d'une prescription par l'entremise du régime ». Les moyennes de domaine sont généralement concentrées vers l'extrémité supérieure des échelles respectives, indiquant que la plupart des réponses sont favorables.

Tableau 1

Distribution des réponses et des évaluations pour des questions de même type (n = 705 848 répondants)

Statistique	Numérique	Fréquence	Problème	Oui/non
Nombre de questions	4	11	11	9
Pourcentage de réponses				
Moyenne	74,97	62,56	30,32	57,26
Minimum	50,90	27,70	4,00	12,50
Maximum	95,00	74,50	64,40	96,70
Moyenne de questions				
Moyenne	8,76	3,57	2,70	1,78
Minimum	8,57	3,09	2,49	1,62
Maximum	8,88	3,84	2,86	1,97
Distribution des évaluations (entre les questions de même type)				
0	0,5			
1	0,4	2,0	5,7	19,5
2	0,4	6,3	12,1	80,5
3	0,7	23,9	82,2	
4	0,9	67,8		
5	4,6			
6	3,0			
7	6,2			
8	16,1			
9	17,8			
10	49,5			

Les questions comportent une échelle numérique 0-10 allant de « pire possible » à « meilleur possible », une échelle de « fréquence » ou une échelle d'« intensité de problème » ordonnée de trois points 1-3 (pas un problème/plus ou moins un problème/un gros problème) ou sont dichotomiques 1-2 (non/oui).

Nous présentons aussi dans le tableau 1 la moyenne, le minimum et le maximum de domaine pour l'ensemble des questions de même type. Ces données montrent que les questions à échelle 0-10 sont celles pour lesquelles la variation totale est la plus faible (après rééchantonnement pour produire la fourchette 0-1 commune), tandis que les questions à échelle 1-2 sont celles dont la variation totale est la plus importante entre les domaines et entre les questions. Ces observations se dégagent aussi de la figure 1, où nous observons que la distribution des questions à échelle 1-2 varie considérablement d'une question à l'autre, tandis que celle des questions à échelle 0-10 est plus homogène.

Le tableau 2 donne les statistiques sommaires pour les moyennes et les écarts-types des évaluations moyennes de domaine, calculées sur l'ensemble des questions de même type. Ces données complètent celles de la figure 1 en résumant les différences entre les distributions des questions

domaine h). L'estimateur résultant pour le domaine h (pour les variances et les corrélations transformées) est :

$$\hat{f}_h = \frac{\hat{\sigma}_h^2 \hat{f}_{dr}^h + \sigma_h^2 \hat{f}_{mod}^h}{\hat{\sigma}_h^2 + \sigma_h^2} = \hat{f}_{dr}^h + \frac{\sigma_h^2}{\hat{\sigma}_h^2 + \sigma_h^2} (\hat{f}_{mod}^h - \hat{f}_{dr}^h),$$

où \hat{f}_{dr}^h et \hat{f}_{mod}^h représentent les estimateurs directs et fondés sur un modèle. Cette formule générale s'applique aux estimations de la variance pour l'ensemble des questions et aux estimations des corrélations pour toutes les paires de questions. L'expression la plus à droite a la forme d'un estimateur bayésien empirique.

Si l'estimateur de variance direct et celui fondé sur un modèle sont indépendants, la variance de l'estimateur combiné résultant est $\tau^2 \sigma_h^2 / (\tau^2 + \sigma_h^2) \leq \min\{\tau^2, \sigma_h^2\}$. Donc, l'estimateur combiné est, au moins, aussi précis que l'un ou l'autre des deux estimateurs qui le constituent, donc entre la prédiction directe et celle fondée sur un modèle. Cette stratégie est utile, surtout quand les prédictions fondées sur un modèle sont meilleures que les estimations directes pour certains domaines, mais non tous.

4. Exemple : Ensemble de données de la CAHPS®

La Consumer Assessments of Health Plans Study (CAHPS®) (Goldstein, Cleary, Langwell, Zaslavsky et Heller, 2001) a été conçue principalement pour recueillir les évaluations et les déclarations des consommateurs au sujet des régimes d'assurance-maladie. Les scores moyens des régimes (peuvent-être après recodage) pour les diverses questions du sondage sont calculés et communiqués aux consommateurs, aux régimes d'assurance-maladie et aux acheteurs. Chaque domaine d'analyse comprend les personnes inscrites à un régime d'assurance-maladie (ou une partie géographique définie d'un régime) durant une année particulière; la plupart des régimes sont échantillonnés pour plusieurs années. La strate est l'unité déclarante (régime ou partie de celui-ci) durant une année donnée; les unités déclarantes correspondent à des régimes sauf dans le cas de quelques régimes très importants comptant plusieurs unités déclarantes. Par conséquent, le nombre d'unités pour l'estimation des fonctions de variance et de covariance est grand.

Nous illustrons notre méthode au moyen d'un ensemble de données de la CAHPS pour les bénéficiaires des régimes de soins américains gérés par Medicare, c'est-à-dire un système d'entités privées, mais financées par les deniers publics qui ont desservi de 5,7 à 6,9 millions de bénéficiaires âgés ou handicapés chaque année de la période couverte par l'étude (1997 à 2001). Nos données représentent 381 domaines déclarants, chacun échantillonné dans

4.1 Statistiques descriptives

Le tableau 1 présente la répartition des réponses et celle des moyennes de domaine selon le type de question. Les observations manquant à cause d'un enchaînement de questions structurées surviennent souvent en blocs, le nombre de questions sautées pouvant aller jusqu'à 11 sur la base d'une seule question filre. La proportion de non-réponses n'étant pas due à un enchaînement de questions structurées est très faible (moins de 2 % pour presque toutes les questions). Dans la présente analyse, nous traitons tous les types de non-réponses de la même façon.

Les taux les plus faibles de réponse à une question (aussi faibles que 4 %) sont ceux enregistrés pour les questions à échelle d'intensité de problème, dont plusieurs ont trait à des services spécialisés, comme des traitements ou des soins de santé à domicile dont ont besoin un assez petit nombre de répondants seulement. Certaines questions à échelle de fréquence et à réponse oui/non produisent des taux de

égales et des approximations sont notées pour le cas de l'échantillonnage avec probabilités inégales. De façon générale, les estimateurs directs \tilde{f}_h , les valeurs réelles f_h et les prédictions de modèle \tilde{f}_h sont reliés au moyen du modèle hiérarchique

$$\text{Niveau I : } \tilde{f}_h = f_h + e_h, \quad (17)$$

$$\text{Niveau II : } f_h = \tilde{f}_h + e_h, \quad (18)$$

où $e_h \sim [0, \sigma_e^2/R_{s_h}]$, $e_h \sim [0, \tau_e^2]$, et $[e_h, \sigma_e^2]$ indiquent une loi d'espérance μ et de variance σ^2 , mais de forme non

précisée. Dans le cas de l'échantillonnage avec probabilités inégales, nous remplaçons R_{s_h} par $R_{s_h}^*$. Ici, e_h représente l'erreur d'échantillonnage et e_h , l'erreur de modélisation. Marginalement, $\tilde{f}_h = f_h + e_h + e_h$, de sorte que, dans la régression, nous pondérons l'observation pour le domaine h par $w_h = (\tau_e^2 + \sigma_e^2/R_{s_h})^{-1}$, qui est l'inverse de la variance marginale. Sous échantillonnage avec probabilités égales, la variance de l'estimation directe de $\sigma_h^2 = E[f_h - f_h]^2$ est

donnée par

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{1}{\left\{ \frac{R_{s_h} - 1}{R_{s_h}} \sum_{k \in S} (\tilde{y}_{h,k} - M_{h, \tilde{f}_h, k})^2 - (1 - \frac{R_{s_h}}{3}) \tilde{f}_h^2 \right\}}$$

si f est une variance

(19)

et par

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{R_{s_h} - 3}{4} \text{ si } f \text{ est une corrélation transformée. (20)}$$

Dans le cas de l'échantillonnage avec probabilités égales, l'équation (19) est exacte et ne dépend pas des hypothèses paramétriques (Seber 1977, page 14). L'approximation asymptotique (20) de la variance de la corrélation transformée Z_h (Fruand et Walpole 1987, page 477) se détériore à mesure que diminuent les tailles d'échantillon et elle échoue entièrement pour $R_{s_h} \leq 3$. Cependant, les domaines pour lesquels les échantillons sont de petite taille ont peu d'effet sur les modèles ajustés; nous excluons donc les domaines pour lesquels $R_{s_h} \leq 3$ de la modélisation des corrélations.

Lorsque les probabilités d'échantillonnage ne sont pas égales, nous pouvons utiliser la grande contrepartie d'échantillon de (19), donnée par

$$\hat{\sigma}_h^2(\tilde{f}_h) = \sum_{k \in S} \left\{ \left(\frac{\sum_{l \in S} \tilde{f}_{h,l}}{2w_h} - M_{h, \tilde{f}_h, k} \right)^2 - \frac{\sum_{l \in S} \tilde{f}_{h,l}^2}{2} \right\} \times \left(\frac{\tilde{y}_{h,k}}{\tilde{f}_{h,k}} - M_{h, \tilde{f}_h, k} \right) - \frac{\sum_{l \in S} \tilde{f}_{h,l}^2}{2}$$

3.5 Estimateurs combinés

Pour les domaines dont l'échantillon est petit, les estimations directes de la variance de sondage sont trop imprécises pour être utiles, tandis que, pour les domaines plus grands dans la même étude, les estimations peuvent être assez fiables. Fay et Herriot (1979), ainsi que Ghosh et Rao (1994) ont montré que réduire les estimations directes vers une valeur lissée fondée sur un modèle peut améliorer considérablement la précision. Ils proposent des estimateurs pondérés des estimateurs directs et d'estimateurs fondés sur un modèle. Autrement dit, au lieu d'utiliser les estimations directes ou celles obtenues par modélisation générale lissée de la variance/covariance, nous utilisons une moyenne pondérée des deux estimateurs pour éventuellement obtenir d'encore meilleures estimations.

Nous pouvons construire de tels estimateurs pondérés pour les variances de domaine en utilisant le modèle spécifique en (17) et (18). Une approche naturelle consiste à appliquer aux estimateurs directs fondés sur un modèle une pondération inversement proportionnelle aux variances d'échantillonnage et d'erreur de modélisation correspondantes, pour le respectivement (notées σ_h^2 et τ_e^2 , respectivement, pour le

La variance de l'erreur de modélisation τ_e^2 est estimée par :

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{R_{s_h} - 3}{4}.$$

leur corrigé pour l'effet de plan.

pas égales, nous proposons de remplacer (20) par l'estimation corrigée pour l'effet de plan.

(par exemple, les modèles avec logarithme des moyennes comme prédicteur) n'améliorent pas considérablement l'ajustement. En dernière analyse, nous retenons la série de modèles emboîtés qui suit.

$$\text{Modèle C1 : } Z_{h,ij} = \alpha_{0ij}, \quad (11)$$

$$\text{Modèle C2 : } Z_{h,ij} = \alpha_{0ij} + \alpha_{3ij} p_{h,i} p_{h,j}, \quad (12)$$

$$\text{Modèle C3 : } Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} (p_{h,i} + p_{h,j}) + \alpha_{3ij} p_{h,i} p_{h,j}, \quad (13)$$

$$\text{Modèle C4 : } Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij} p_{h,i} p_{h,j}, \quad (14)$$

$$\text{Modèle C5 : } Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij} p_{h,i} p_{h,j} + \alpha_{4ij} p_{h,i}^2 + \alpha_{5ij} p_{h,j}^2, \quad (15)$$

Le modèle C3 est le modèle C4 avec la contrainte $\alpha_{1ij} = \alpha_{2ij}$.

3.3 Prédiction des covariances avec données manquantes structurées

Lorsque les données comportent des enchaînements de

questions, les corrélations d'échantillon des évaluations pour l'ensemble des répondants qui ont répondu aux deux questions peuvent être modélisées au moyen des modèles (11) à (15), comme dans le cas de la réponse complète. Il est facile d'estimer les covariances d'échantillon correspondantes en utilisant les fonctions de variance ajustées pour rééchantillonner les corrélations prélevées. Cependant, comme la covariance d'échantillon reflète la variabilité dans l'ensemble du processus d'échantillonnage, et non simplement la variabilité dans la sous-population de répondants qui ont répondu aux deux questions, la relation entre la covariance d'échantillon et la covariance d'échantillonnage est plus complexe que si les données étaient complètes. À la présente section, nous dérivons la relation entre la covariance d'échantillon pour l'ensemble de répondants qui ont répondu aux deux questions et la covariance d'échantillonnage. Cela nous permet d'appliquer des modèles de corrélation tels que (11) à (15) à des données avec enchaînements de questions.

Pour toute paire de questions, il existe quatre schémas de données, à savoir une réponse aux deux questions, une réponse et une question sautée (deux schémas), et deux questions sautées. Nous étendons notre notation en introduisant un indice supérieur représentant la situation de réponse à une deuxième question. Soit $Y_{h,ij}^1 = \sum_s \bar{y}_{h,ik} \bar{r}_{h,jk} (1 - \bar{r}_{h,jk})$, $R_{h,ij}^1 = \sum_s \bar{r}_{h,ik} \bar{r}_{h,jk}$, $Y_{h,ij}^0 = \sum_s \bar{y}_{h,ik} \bar{r}_{h,jk}$, $R_{h,ij}^0 = \sum_s \bar{r}_{h,ik} \bar{r}_{h,jk}$.

Dans le cas de l'échantillonnage avec probabilités égales, le remplacement de $M_{h,i}$ par l'expression susmentionnée dans (7) donne

$$\bar{V}_{h,ij} = \frac{\bar{R}_{h,ij}^1 \bar{R}_{h,j}^1}{\bar{R}_{h,ij}^0 \bar{D}_{h,ij} \bar{R}_{h,i}^0 \bar{D}_{h,i}^1} \left\{ \bar{C}_{h,ij}^1 + \frac{\bar{R}_{h,ij}^1 \bar{R}_{h,j}^1}{\bar{R}_{h,ij}^0 \bar{D}_{h,ij} \bar{R}_{h,i}^0 \bar{D}_{h,i}^1} \right\}, \quad (16)$$

où $\bar{D}_{h,ij} = \bar{M}_{h,ij}^1 - \bar{M}_{h,ij}^0$. Ici, $\bar{C}_{h,ij}^1 = \sum_s (\bar{y}_{h,ik} - \bar{M}_{h,ik}^1)(\bar{y}_{h,jk} - \bar{M}_{h,jk}^1) / \bar{R}_{h,ij}^1$ est la covariance d'échantillon normalisée des évaluations pour l'ensemble des répondants qui ont répondu aux deux questions (que l'on peut prédire en utilisant les fonctions de corrélation et de variance et, dans le cas de l'échantillonnage avec probabilités inégales, en appliquant un facteur de normalisation). Lorsque les probabilités d'échantillonnage ne sont pas égales, l'équation (16) n'est vérifiée exactement que si $\sum_s \bar{r}_{h,ik} \bar{r}_{h,jk} - \bar{M}_{h,ik}^1 \bar{M}_{h,jk}^1 = 0$. Par conséquent, nous pouvons nous attendre à ce que (16) donne une bonne approximation si les probabilités d'échantillonnage pour une question ne sont pas fortement corrélées aux résidus d'une autre question. En général, il convient de vérifier s'il est approprié d'utiliser (16) pour les plans d'échantillonnage avec probabilités inégales.

Les différences estimées entre les moyennes $\bar{D}_{h,ij}$ déterminent la contribution du schéma de réponse à la covariance d'échantillonnage. Nous pouvons modéliser $\bar{D}_{h,ij}$ ou $\bar{D}_{h,ji}$ dans le processus d'obtention d'estimations issues de $\bar{V}_{h,ij}$. Dans notre application, les $\bar{D}_{h,ij}$ sont généralement petites. Comme le deuxième terme de (16) est un produit de deux facteurs de petite taille ($\bar{D}_{h,ij}$ et $\bar{D}_{h,ji}$), la contribution de $\bar{D}_{h,ij}$ à (16) est faible et il suffit d'utiliser un modèle simple pour $\bar{D}_{h,ij}$, comme une constante pour chaque paire de questions. Cependant, une constante propre à chaque paire de questions. Cependant, une constante propre à chaque paire de questions.

3.4 Ajustement et évaluation des modèles

Nous estimons les paramètres de la fonction de variance ou de la fonction de corrélation par régression par les moindres carrés répondée itérativement. La pondération est importante quand le nombre de réponses varie considérablement d'un domaine à l'autre, comme dans le cas de notre exemple.

À la présente section, nous utilisons un indice pour les domaines (h) et pour les répondants (k), mais non pour les questions, car la même méthodologie s'applique à chaque fait pour le cas de l'échantillonnage avec probabilités

3. Modèles pour les fonctions de variance

À la présente section, nous proposons des spécifications de modèle pour les variances et pour les corrélations d'échantillon pour les réponses complètes ou pour celles

avec enchaînements de questions structurées. Puis, nous discutons des stratégies d'ajustement et d'évaluation des modèles. Nous supposons que ces domaines sont des strates non chevauchantes, de sorte que les erreurs d'échantillonnage pour divers domaines soient indépendantes.

Nous transformons les évaluations ordonnées pour les

amener à l'intervalle $[0, 1]$ par la transformation $p_{h,i} =$

$(B_{h,i} - M_{h,i}) / (B_{h,i} - A_{h,i})$, où $A_{h,i}$ et $B_{h,i}$ sont les catégories de réponse minimale et maximale pour la question i

dans le domaine h , respectivement. Nous nous concentrons

sur la modélisation des variances pour les grandes valeurs de $M_{h,i}$ (petites valeurs de $p_{h,i}$) parce que, dans l'exemple

que nous avons choisi, les résultats moyens sont habituellement proches de l'extrémité supérieure de l'échelle.

3.1 Fonctions de variance

Afin de tenir compte du nombre variable de répondants sur les domaines et les questions, et des différences

d'échelle, nous normalisons les estimateurs de variance donnés par (6) pour la taille d'échantillon et faisons un

rééchantillonnage :

$$\tilde{V}_{h,i}^{h,i} = \frac{\tilde{R}_{s_{h,i}} \tilde{V}_{h,i}^{h,i}}{(B_{h,i} - A_{h,i})^2}.$$

Sous échantillonnage avec probabilités inégales dans les domaines, nous pourrions utiliser un facteur de normalisation qui tient compte des pondérations. Une normalisation

possible consiste à multiplier $\tilde{V}_{h,i}^{h,i}$ par $\tilde{R}_{s_{h,i}}^2 / (\sum \tilde{r}_{h,i,k})^2$, où $\tilde{r}_{h,i,k}$ est l'indicateur de réponse à la question i

pour le k^e sujet dans le h^e domaine, à la place de $\tilde{R}_{s_{h,i}}$. Cette

approximation, proposée par Kish (1965), possède une justification fondée sur un modèle (Gabler, Haeder et Lahiri 1999). Elle donne de bons résultats si les probabilités

d'échantillonnage varient moyennement dans l'échantillon, mais peut être inefficace si la variation est excessive (Kom

et Graubard 1999, page 173; Spencer 2000).

Comme, dans notre exemple, les questions ont des valeurs ordonnées, la variance doit tendre vers 0 quand

$p_{h,i} \rightarrow 0$ ou $p_{h,i} \rightarrow 1$. Un prédicteur ayant manifestement

cette propriété est la fonction de variance de la loi de Bernoulli, $p_{h,i}(1 - p_{h,i})$. Celle-ci est vérifiée exactement

pour les questions dichotomiques et pourrait être une approximation utile pour les questions comportant au moins

trois catégories de réponse.

Comme autres solutions que le modèle de variance de

Bernoulli, nous considérons des modèles contenant diverses fonctions polynomiales et autres des moyennes comme

prédicteur. De tous les modèles envisagés, la famille de modèles quadratiques a donné des résultats d'ajustement aussi bons que n'importe quelle autre. Nous nous concentrons sur les modèles quadratiques qui suivent.

$$\text{Modèle V1 : } \tilde{V}_{h,i}^{h,i} = \beta_{11} p_{h,i}, \quad (8)$$

$$\text{Modèle V2 : } \tilde{V}_{h,i}^{h,i} = \beta_{21} p_{h,i} (1 - p_{h,i}), \quad (9)$$

$$\text{Modèle V3 : } \tilde{V}_{h,i}^{h,i} = \beta_{11} p_{h,i} + \beta_{21} p_{h,i} (1 - p_{h,i}). \quad (10)$$

Donc, nous considérons un modèle de variance linéaire V1, un modèle de type binomial V2 et un modèle de variance quadratique général V3. Tous ces modèles assurent correctement que $\tilde{V}_{h,i}^{h,i} = 0$ quand $p_{h,i} = 0$, mais seul V2 assure que $\tilde{V}_{h,i}^{h,i} = 0$ quand $p_{h,i} = 1$. La logique qui sous-tend V1 est que les relations sont souvent approximativement linéaires sur de petits intervalles. Aussi bien V1 que V2 sont des sous-modèles du modèle quadratique à deux paramètres V3. Nous avons également considéré des modèles pour $\log(\tilde{V}_{h,i}^{h,i})$, mais ceux-ci n'ont pas donné un

aussi bon ajustement.

Le modèle V3 est équivalent au modèle proposé par Wolter (1985, chapitre 5) : l'équivalence se voit en

exprimant le deuxième membre de V3 en fonction de $p_{h,i}$ et de $p_{h,i}^2$, puis en divisant les deux membres par $p_{h,i}^2$ pour obtenir la variance relative. Cependant, les estimations des

paramètres obtenues par ajustement des deux formes du

modèle peuvent différer selon les hypothèses de modélisation utilisées.

3.2 Fonctions de corrélation avec données complètes

Comme les corrélations sont indépendantes de l'échelle des données, nous les modélisons et nous dérivons les covariances d'échantillonnage, au lieu de modéliser directement les covariances. Nous modélisons les corrélations

d'échantillon

$$\rho_{h,i,j} = \frac{\tilde{V}_{h,i,j}^{h,i,j}}{\tilde{V}_{h,i,i}^{h,i,i} \tilde{V}_{h,j,j}^{h,j,j}}^{1/2},$$

par la voie des valeurs transformées non contraintes $Z_{h,i,j} = \log\{(1 + \rho_{h,i,j}) / (1 - \rho_{h,i,j})\}$. Contrairement aux modèles de variance, les modèles de corrélation peuvent inclure une ordonnée à l'origine non contrainte, puisque la corrélation n'est sujette à aucune contrainte naturelle quand $p_{h,i}$ ou $p_{h,j}$ s'approche de 0 ou de 1.

Puisque $\rho_{h,i,j}$ est une fonction des premier et deuxième moments des questions i et j , il semble raisonnable de se concentrer d'abord sur les modèles linéaires et quadratiques pour $Z_{h,i,j}$. Comme pour les fonctions de variance, nous constatons qu'une gamme plus étendue de modèles

l'échantillon, respectivement, pour le h^e domaine, $Y_{h,i} = \sum_{U_h} Y_{h,ik}$, $R_{h,i} = \sum_{U_h} R_{h,ik}$, $\bar{Y}_{h,i} = \sum_{S_h} \bar{Y}_{h,ik}$ et $\bar{R}_{h,i} = \sum_{S_h} \bar{R}_{h,ik}$, où $\bar{Y}_{h,ik} = y_{h,ik} / \pi_{h,k}$, $\bar{R}_{h,ik} = r_{h,ik} / \pi_{h,k}$, et $\pi_{h,k} = \text{pr}(k \in S_h)$.

Le vecteur des résultats moyens pour la population d'éléments compris dans le domaine h est

$$M_h = f(Y_h, R_h) = \left(\frac{Y_{h,1}}{R_{h,1}}, \dots, \frac{Y_{h,l}}{R_{h,l}} \right),$$

où $Y_h = (Y_{h,1}, \dots, Y_{h,l})$ et $R_h = (R_{h,1}, \dots, R_{h,l})$. Un estimateur est donné par

$$f(\hat{Y}_h, \hat{R}_h) = \left(\frac{\hat{Y}_{h,1}}{\hat{R}_{h,1}}, \dots, \frac{\hat{Y}_{h,l}}{\hat{R}_{h,l}} \right).$$

Un développement en série de Taylor de premier ordre de $f(\hat{Y}_h, \hat{R}_h)$ autour de $f(Y_h, R_h)$ produit l'approximation

$$\text{var}(f(\hat{Y}_h, \hat{R}_h)) \approx V_h = f'(Y_h, R_h) \text{var}(\hat{Y}_h, \hat{R}_h) f'(Y_h, R_h)^T,$$

où $f'(Y_h, R_h)$ est le jacobien de $f(Y_h, R_h)$. Souvent, il est informaticquement plus facile de commencer par calculer $u_{h,k} = f'(Y_h, R_h)^T z_{h,k}$, où $z_{h,k} = (y_{h,k}, r_{h,k})$, puis d'évaluer la variance sous la forme

$$V_h = \text{var} \left(\sum_{S_h} \bar{u}_{h,k} \right) = \text{var} \left(\sum_{U_h} \bar{u}_{h,k} I_{h,k} \right) = \sum_{k,l \in U_h} \Delta_{h,kl} \bar{u}_{h,k} \bar{u}_{h,l}^T,$$

où $I_{h,k} = 1$ si $k \in S_h$ (indiquant que le k^e membre du domaine h est échantillonné) et 0 autrement, $\Delta_{h,kl} = \pi_{h,kl} - \pi_{h,k} \pi_{h,l}$, et $\pi_{h,kl} = \text{pr}(k, l \in S_h)$. Un estimateur de V_h est

$$\hat{V}_h = \sum_{k,l \in S_h} \bar{\Delta}_{h,kl} \bar{u}_{h,k} \bar{u}_{h,l}^T, \quad (1)$$

où $\bar{\Delta}_{h,kl} = \Delta_{h,kl} / \pi_{h,kl}$.

Pour décrire l'évaluation de \hat{V}_h , nous ne devons considérer qu'un seul élément diagonal (c 'est-à-dire, variance) et un seul élément hors diagonale (c 'est-à-dire, covariance). La sous-matrice du jacobien formée par les i^e et j^e questions est

$$f'(Y_h, R_h) = \begin{pmatrix} \frac{1}{R_{h,i}} & 0 & \frac{R_{h,j}}{1} & 0 \\ \frac{R_{h,i}}{1} & 0 & 0 & -\frac{R_{h,i}^2}{Y_{h,i}} \\ 0 & -\frac{R_{h,i}^2}{Y_{h,i}} & 0 & -\frac{R_{h,j}^2}{Y_{h,j}} \\ 0 & 0 & 0 & -\frac{R_{h,j}^2}{Y_{h,j}} \end{pmatrix}.$$

Pour $z_{h,k} = (y_{h,ik}, y_{h,jk}, r_{h,ik}, r_{h,jk})$, il s'ensuit que

$$\hat{V}_{h,ii} = \frac{R_{h,i}^2}{1} \sum_{k,l \in S_h} \bar{\Delta}_{h,kl} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik})(\bar{y}_{h,il} - M_{h,i} \bar{r}_{h,il}) \quad (2)$$

où $M_{h,i} = Y_{h,i} / R_{h,i}$ est le résultat moyen de la i^e question dans le domaine h . Donc,

$$u_{h,k} = f'(Y_h, R_h) z_{h,k} = \begin{pmatrix} \frac{1}{R_{h,i}} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik}) \\ \frac{1}{R_{h,i}} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik}) \\ \frac{1}{R_{h,j}} (\bar{y}_{h,jk} - M_{h,j} \bar{r}_{h,jk}) \\ \frac{1}{R_{h,j}} (\bar{y}_{h,jk} - M_{h,j} \bar{r}_{h,jk}) \end{pmatrix},$$

et

$$\hat{V}_{h,ij} = \frac{1}{1} \sum_{k,l \in S_h} \bar{\Delta}_{h,kl} (R_{h,i} \bar{r}_{h,ik} - M_{h,i} \bar{r}_{h,ik})(\bar{y}_{h,il} - M_{h,i} \bar{r}_{h,il}) \times (\bar{y}_{h,jl} - M_{h,j} \bar{r}_{h,jl}). \quad (3)$$

Pour évaluer (2) et (3), nous faisons une autre approximation en substituant $\bar{R}_{h,i} = \sum_{S_h} \bar{r}_{h,ik}$ et $\bar{M}_{h,i} = \sum_{S_h} \bar{y}_{h,ik} / (\sum_{S_h} \bar{r}_{h,ik})$ à $R_{h,i}$ et $M_{h,i}$.

Si les taux d'échantillonnage sont faibles ou que nous souhitions faire des prédictions pour une grande super-population (par exemple, tous les participants possibles à un régime d'assurance-maladie plutôt que ceux couramment inscrits seulement), $\bar{\Delta}_{h,kl} = 1 - \pi_{h,k} \approx 1$ si $k = l$, $\bar{\Delta}_{h,kl} = 0$ si $k \neq l$, et le plan d'échantillonnage s'approche de l'échantillonnage avec remise. Sous le plan d'échantillonnage avec remise, les estimateurs approximativement sans biais sont

$$\hat{V}_{h,ii} = \frac{R_{h,i}^2}{1} \sum_{k \in S_h} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik})^2 \quad (4)$$

et

$$\hat{V}_{h,ij} = \frac{1}{1} \sum_{k \in S_h} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik})(\bar{y}_{h,jk} - M_{h,j} \bar{r}_{h,jk}). \quad (5)$$

Ces estimateurs peuvent être généralisés pour prendre en compte la mise en grappes. En cas d'échantillonnage avec probabilités égales dans les domaines, (4) et (5) se réduisent à

$$\hat{V}_{h,ii} = \frac{R_{h,i}^2}{1} \sum_{k \in S_h} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik})^2 \quad (6)$$

et

$$\hat{V}_{h,ij} = \frac{1}{1} \sum_{k \in S_h} (\bar{y}_{h,ik} - M_{h,i} \bar{r}_{h,ik})(\bar{y}_{h,jk} - M_{h,j} \bar{r}_{h,jk}), \quad (7)$$

où $\bar{R}_{h,i}$ est le nombre de répondants à la question i dans le domaine h .

choix naturel. D'autres FVG proposées ont également une forme simple (Woodruff 1992; Otto et Bell 1995).

Trouver une FVG appropriée peut simplifier les calculs et rendre les estimations de la variance plus stables. En outre, résumer les estimations de la variance d'échantillon-nage sous la forme d'une fonction facilite la présentation de grandes quantités de statistiques (Wolter 1985, pages 201-202). Enfin, modéliser les variances sous forme de fonctions des moyennes facilite la réestimation itérative des variances d'échantillon-nage en cas de modélisation hiérarchique. En pratique, la décision d'utiliser des fonctions de variance dans un contexte de modélisation hiérarchique dépend de la qualité de l'ajustement de la FVG; l'utilisation de cette dernière ne vaut la peine que si l'ajustement est suffisamment bon.

Les études antérieures sur les FVG sont assez rares. Wolter (1985, chapitre 5) donne un aperçu, mais ne fournit que quelques références, comme le font aussi Valliant, Dorfman et Royall (2000, pages 344 à 348). Valliant (1992a, 1992b) utilise des FVG pour lisser des indices variables en fonction du temps dans les analyses de séries chronologiques. Woodruff (1992) utilise des FVG pour estimer la variance de la variation de l'emploi dans la Current Employment Survey, et Wolter (1985, pages 208 à 217) illustre l'utilisation des FVG au moyen de données provenant de la Current Population Survey. Des FVG sont également utilisées dans la National Health Interview Survey (Valliant et coll. 2000, page 344).

Huff, Eltinge et Gershunskaya (2002), ainsi que Cho, Eltinge, Gershunskaya et Huff (2002) considèrent l'utilisation de FVG pour la Current Employment Survey et la Consumer Expenditure Survey réalisées aux États-Unis. Eltinge (2002) utilise des FVG pour estimer une matrice des covariances d'échantillon-nage complète, lorsque les échantillons sont trop petits pour produire des estimations stables pour toutes les régions, et estime les composantes de l'erreur quadratique moyenne (MSE pour *mean squared error*) du modèle de FVG. Otto et Bell (1995) ajustent des FVG au revenu médian, au revenu par habitant et au taux de pauvreté selon le groupe d'âge dans la Current Population Survey, en supposant que l'interdépendance des taux au cours du temps est autorégressive et que les matrices des covariances d'échantillon-nage suivent une loi de Wishart.

Notre étude prolonge les travaux antérieurs sur les FVG dans quatre directions. En premier lieu, nous utilisons la FVG pour faire une généralisation sur l'ensemble des domaines plutôt que sur l'ensemble des questions. Donc, nous ne supposons pas que les diverses questions ont la même forme pour les questions ayant des catégories de réponses semblables. En deuxième lieu, nous élaborons des FVG pour la matrice des covariances

complète, qui doivent être estimées pour une inférence conjointe sur plusieurs résultats. En troisième lieu, nous nous concentrons sur la relation entre les moyennes et les variances des questions à format de réponse ordonné souvent utilisées dans les questionnaires d'enquête, plutôt que sur des réponses continues homoscédastiques. Enfin, nous tenons compte explicitement des profils de non-réponse due à des enchaînements de questions structurées. Alors qu'on peut ignorer la non-réponse partielle structurée (sauf son effet sur la taille d'échantillon) dans le cas de l'estimation univariée, il faut en tenir compte explicitement pour modéliser les relations bivariées, parce qu'elle a une incidence sur la covariance d'échantillon-nage des moyennes des questions. De surcroît, comme le nombre de réponses varie selon la question, nous ne pouvons modéliser les covariances d'échantillon-nage au moyen de la loi de Wishart, qui ne possède qu'un seul paramètre pour la taille d'échantillon.

Nous commençons par décrire l'estimation directe des variances et des covariances, y compris le cas où des données manquent à cause d'enchaînements de questions. À la section 3, nous présentons des modèles pour les fonctions de variance et de covariance généralisées (FVCG) et nous exposons nos stratégies d'ajustement et d'évaluation de modèles, et de combinaison d'estimations directes et de prédictions par modèle. À la section 4, nous appliquons nos méthodes à une grande enquête sur les soins de santé. À la section 5, pour conclure, nous décrivons des applications et des extensions de nos méthodes.

2. Estimations directes des variances d'échantillon-nage des moyennes de domaine

Nous indiquons les observations par domaine (indice h), par question (indices i et j) et par répondant (indices k et l); $y_{h,ik}$ et $r_{h,ik}$ représentent le résultat et l'indicateur de réponse du sujet k dans le domaine h pour la question i . Nous supprimons l'indice inférieur de question quand nous faisons référence à l'ensemble des questions pour un répondant ou un domaine, et nous n'avons pas besoin d'utiliser d'indice inférieur de répondant quand nous discutons des moyennes, des variances et des corrélations de questions. L'estimation directe de la matrice des covariances d'échantillon-nage des moyennes de domaine (donc, « estimation de variance ») débute par l'expression des moyennes sous forme de fonctions des totaux des résultats et des indicateurs de réponse. Nous remplaçons $y_{h,ik}$ par 0 pour les observations manquantes, de sorte que les totaux soient définis en présence d'enchaînements de questions. Suivant la notation de Samdal, Swensdal, et Wirtman (1992, pages 24 à 28; 36 à 42), soit U_h et S_h la population et

Fonctions de variance-covariance pour les moyennes de domaine des questions avec valeurs ordonnées

Alister James O'Malley et Alan Mark Zaslavsky

Résumé

De nombreuses analyses statistiques, particulièrement l'analyse multivariée, requièrent l'estimation d'une matrice des variances-covariances d'échantillonnage. Dans le cas de problèmes univariés, des fonctions reliant la variance à la moyenne ont été utilisées pour obtenir des estimations de la variance, en regroupant l'information sur l'ensemble des unités ou des variables. Nous présentons des fonctions de variance et de corrélation pour des moyennes multivariées de questions d'enquête avec valeurs ordonnées, pour des données complètes, ainsi que pour des données avec non-réponses structurées. Nous élaborons aussi des méthodes permettant d'évaluer l'ajustement du modèle et de calculer des estimateurs composites qui combinent des prédictions directes et fondées sur un modèle. Nous utilisons des données d'enquête provenant de la Consumer Assessments of Health Plans Study (CAHPS®) pour illustrer l'application de la méthodologie.

Mots clés : Fonction de variance; fonction de corrélation; modèle hiérarchique; réponse ordonnée, non-réponse; enchaînement de questions.

1. Introduction

Les données d'enquête sont souvent utilisées pour obtenir des mesures permettant de faire des comparaisons entre domaines d'estimation. Dans notre exemple pratique, des enquêtes sont réalisées pour recueillir des déclarations sur les expériences vécues en ce qui concerne les régimes d'assurance-maladie (entités qui administrent les soins de santé) auprès des membres inscrits; de même, une enquête pourrait être conçue pour évaluer les écoles en faisant passer des tests à un échantillon d'élèves.

Une part essentielle de l'analyse des données d'enquête est le calcul des variances d'échantillonnage ou de la matrice des covariances d'échantillonnage d'un estimateur multivarié. L'approche type en échantillonnage consiste à calculer les variances directement pour chaque estimateur dans chaque domaine. Les estimations directes de la variance peuvent être instables si le nombre de répondants à une question est faible parce que la taille de l'échantillon pour un domaine est petite, vu que la question s'applique seulement à une fraction des répondants (comme les utilisateurs d'équipement spécialisé dans les enquêtes sur la santé), ou parce que nous soustrayons les moyennes pour un petit sous-groupe (comme les personnes atteintes de maladie chronique).

En modélisant les estimations de la variance sous forme de fonctions des moyennes d'unité (domaine), nous pouvons regrouper l'information sur l'ensemble des unités pour obtenir des estimations plus stables. Bien que la modélisation puisse introduire un biais, pour les petites unités, ce problème est compensé par la réduction de la variation

d'échantillonnage. On peut aussi envisager de généraliser les estimations de la variance sur l'ensemble des questions en plus des domaines, ou à la place de ceux-ci. Cette approche convient lorsqu'il existe des groupes de questions pour lesquelles il est probable que la même relation moyenne-variance soit vérifiée. Cependant, si le nombre de domaines est beaucoup plus grand que celui des questions, l'amélioration éventuelle la plus importante s'obtient en généralisant sur l'ensemble des domaines plutôt que sur l'ensemble des questions.

Une *fonction de variance généralisée* (FVG) est un modèle mathématique qui décrit la relation entre la variance ou la variance relative d'un estimateur et son espérance. Si plusieurs estimations sont produites d'après le même échantillon, Wolter (1985, chapitre 5) propose le modèle

$$V / M^2 = \theta_0 + \theta_1 / M,$$

où M et V représentent la valeur prévue et la variance de l'estimateur, respectivement. Une forme de ce genre pourrait convenir pour des variables, comme le revenu ou la richesse, pour lesquelles un coefficient de variation quasi constant serait plausible, parce que la moyenne et l'écart-type sont proportionnels à la longueur de la période de référence. La modélisation du coefficient de variation est donc pertinente surtout dans les situations où les variables ont un contenu semblable, mais des échelles différentes (par exemple, données sur le revenu recueillies mensuellement et annuellement). Dans notre problème, les questions ont des valeurs ordonnées, si bien qu'un modèle du coefficient de variation n'est pas un

- Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.
- Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Varivartan, S., et Little, R.J.A. (2002). On the formation of weighting adjustment cells for unit nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Remerciements

Remerciements

Bibliographie

- Czajka, J.T., Hirabayashi, S.M., Little, R.J.A. et Rubim, D.B. (1987). Evaluation of a new procedure for estimating income aggregates from advance data. Dans *Statistics of Income and Related Administrative Record Research: 1986-1987*, U.S. Department of the Treasury, 109-136.
- Elliot, M.R., et Little, R.J.A. (2000). Model-based alternatives to weighting survey weights. *Journal of Official Statistics*, 16, 191-209.
- Ezzai, T., et Khare, M. (1992). Nonresponse adjustments in a National Health Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 339-344.
- Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47, 663-685.
- Holt, D., et Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical A*, 142, 33-46.
- Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquêtes manquantes. *Techniques d'enquête*, 12, 1-16.
- Kish, L. (1992). Weighting for unequal P. *Journal of Official Statistics*, 8, 183-200.
- Little, R.J.A. (1986). Survey nonresponse adjustments. *Revue Internationale de la Statistique*, 54, 139-157.
- Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- Little, R.J.A., Lewitzky, S., Heenenga, S., Lepkowski, J. et Kessler, R.C. (1997). An assessment of weighting methodology for the national comorbidity study. *American Journal of Epidemiology*, 146, 439-449.
- Little, R.J.A., et Rubim, D.B. (2002). *Statistical Analysis with Missing Data*, 2^{ème} édition, New York: John Wiley & Sons, Inc.
- Oh, H.T., et Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, 2. Theory and Bibliographies, (Eds. W.G. Madow, I. Olkin et D.B. Rubin), Academic Press, New York, 143-184.

Les résultats des sections 2 et 3 ont des incidences importantes en ce qui concerne l'utilisation de la pondération comme outil de correction pour la non-réponse totale. Les enquêtes comportent souvent de nombreuses variables de résultat auxquelles est habituellement appliquée la même pondération. L'analyse de la section 2 et les simulations de la section 3 donnent à penser que l'on pourrait obtenir de meilleurs résultats en estimant l'erreur quadratique moyenne des moyennes pondérées et non pondérées, et en limitant la pondération aux cas pour lesquels cette relation est importante. Une approche plus perfectionnée consiste à appliquer des modèles à effets aléatoires pour réduire les coefficients de pondération, de telle façon que la réduction soit plus importante pour les résultats qui ne sont pas fortement corrélés aux covariables (par exemple, Elliott et Little (2000)). Une autre option simple est l'imputation fondée sur les modèles de prédiction, puis sur ces derniers pour utiliser des prédicteurs échelonnés par intervalles ainsi que des prédicteurs catégoriques, et de laisser tomber les interactions afin d'intégrer un plus grand nombre d'effets principaux. L'imputation multiple (Rubin (1987)) peut être utilisée pour propager l'incertitude.

Quant on dispose de beaucoup d'information sur les covariables, une façon intéressante d'aborder la généralisation à des corrections par catégorie de pondération consiste à créer un score de propension pour chaque répondant d'après une régression logistique de l'indicateur de non-réponse sur les covariables, puis de créer des cellules d'ajustement de ce score. Les méthodes axées sur le score d'ajustement ont été élaborées au départ dans le contexte des cas appariés et des témoins dans les études par observation (Rosenbaum et Rubin 1983), mais sont appliquées assez fréquemment aujourd'hui dans le contexte de la non-réponse totale (Little 1986; Czajka, Hirabayashi, Little et Rubin 1992; Ezziati et Khare 1999). Ici, l'analyse laisse entendre que, pour que cette approche soit productive, le score de propension doit être prédicteur des résultats. Vartivarian et Little (2002) considèrent des cellules d'ajustement basées sur la classification conjointe en fonction de la propension à répondre, ainsi que des prédicteurs sommaires des résultats afin de tirer parti des associations résiduelles entre les covariables et le résultat après correction pour tenir compte du score de propension. L'existence que les variables de cellule d'ajustement présentent les résultats étaye cette approche.

L'analyse présentée ici pourrait être étendue de plusieurs

L'analyse présentée ici pourrait être étendue de plusieurs façons. Les termes de deuxième ordre figurant dans l'ex-pression de la variance sont ignorés ici, s'ils étaient inclus, ils pénaliseraient la pondération fondée sur un grand nombre de petites cellules d'autrement. Les corrections pour population finie pourraient être incluses, mais il semble peu probable qu'elles aient une incidence sur les principales conclusions. Il serait intéressant de voir dans quelle mesure

Le tableau 5b donne les résultats pour le taux de réponse de 70 %. Le profil des résultats est fort semblable à celui du tableau 5a. Comme prévu, les différences entre les méthodes sont plus petites, quoiqu'elles demeurent considérables pour nombre de lignes du tableau.

lation.

Tableau 5a

Association avec les cellules		d'ajustement basée sur X		Cellule
		(Y, X)	n	
4	Forté	Forté	400	6 955
	Forté	Forté	2 000	7 024
	Forté	Forté	2 000	7 008
	Forté	Forté	2 000	5 376
4	Moyenne	Moyenne	2 000	5 441
	Moyenne	Moyenne	2 000	3 731
	Moyenne	Moyenne	2 000	3 703
	Moyenne	Moyenne	2 000	3 042
4	Forté	Forté	400	2 864
	Forté	Forté	400	1 148
	Forté	Forté	400	1 113
	Forté	Forté	400	2 995
4	Moyenne	Moyenne	400	2 838
	Moyenne	Moyenne	400	3 006
	Moyenne	Moyenne	400	2 991
	Moyenne	Moyenne	400	2 988
3	Forté	Forté	400	476
	Forté	Forté	400	1 178
	Forté	Forté	400	1 178
	Forté	Forté	400	2 988
3	Moyenne	Moyenne	400	3 703
	Moyenne	Moyenne	400	3 703
	Moyenne	Moyenne	400	3 703
	Moyenne	Moyenne	400	3 042
3	Forté	Forté	400	2 864
	Forté	Forté	400	1 148
	Forté	Forté	400	1 113
	Forté	Forté	400	2 995
3	Moyenne	Moyenne	400	2 838
	Moyenne	Moyenne	400	3 006
	Moyenne	Moyenne	400	2 991
	Moyenne	Moyenne	400	2 988
2	Forté	Forté	400	476
	Forté	Forté	400	1 178
	Forté	Forté	400	1 178
	Forté	Forté	400	2 988
2	Moyenne	Moyenne	400	3 703
	Moyenne	Moyenne	400	3 703
	Moyenne	Moyenne	400	3 703
	Moyenne	Moyenne	400	3 042
1	Forté	Forté	400	2 864
	Forté	Forté	400	1 148
	Forté	Forté	400	1 113
	Forté	Forté	400	2 995
1	Moyenne	Moyenne	400	2 838
	Moyenne	Moyenne	400	3 006
	Moyenne	Moyenne	400	2 991
	Moyenne	Moyenne	400	2 988

Tableau 5b

Résultats sommaires des estimateurs basés sur 1 000 échantillons répétés pour $C = 10$ cellules d'ajustement, limitées aux échantillons répétés pour lesquels $n_{0j} > 0$ pour tout j . Le taux de réponse est de 70 %. Les valeurs sont multipliées par 1 000

[illegible]

404	emp.	400	4692	4.810	4.893	4.860	emp.	1.129	1.192	emp.	889	894	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133	1.129	emp.	129	998	emp.	133
-----	------	-----	------	-------	-------	-------	------	-------	-------	------	-----	-----	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----	-------	------	-----	-----	------	-----

Le tableau 5a donne les résultats des simulations pour un taux de réponse de 52 %. Les lignes sont étiquetées d'après les quatre cellules du tableau 1, où les associations moyenne et forte sont combinées. Pour chaque ligne, la plus faible des REQM pour les moyennes des répondants non pondérée et pondérée figure en caractères gras, indiquant la supériorité de la méthode correspondante.

Les quatre premières lignes du tableau 5a correspondent à la cellule 4 du tableau 1, avec une association moyenne/forte entre Y et X , et une association faible entre M et X . Dans ces cas, le biais de \bar{y}_0 n'est plus important, mais la précision de \bar{y}_w s'est améliorée, particulièrement quand l'association entre Y et X est forte. Il s'agit de cas où la pondération réduit la variance au lieu de l'accroître. Les estimations analytiques de la RMSE et les estimations sur échantillon s'approchent des estimations empiriques de la RMSE, tandis que l'équation de Kish la surestime, comme le prédit la théorie exposée à la section 2.

Les deux lignes suivantes du tableau 5a correspondent à la cellule 2 du tableau 1, où l'association entre Y et X est faible et l'association entre M et X est moyenne ou forte. Dans ces cas, la MSE de \bar{y}_w est plus grande que celle de \bar{y}_0 . Ces cas illustrent des situations où la pondération accroit la variance, sans réduction compensatoire du biais. La dernière ligne correspond à la cellule 1 du tableau 1, avec des associations faibles entre M et X et entre Y et X . La moyenne non pondérée a une REQM plus faible dans ces conditions, mais l'accroissement de la REQM dû à la pondération est négligeable. Pour les trois dernières lignes du tableau 5a, la REQM pour l'équation de Kish est semblable à celle produite au moyen de la formule analytique de la section 2 et aux estimations empiriques fondées sur cette formule, et toutes ces estimations s'approchent de la REQM empirique.

Les deux dernières colonnes du tableau 5a donnent le biais et la REQM empiriques avec la méthode composite pour les simulations des six premières lignes, l'estimateur composite est le même que \bar{y}_w et, par conséquent, détecte et élimine le biais de la moyenne non pondérée. Pour les simulations dans la cellule 1 (dernière ligne), l'estimateur composite donne les mêmes résultats que \bar{y}_w ou \bar{y}_0 , comme il fallait s'y attendre puisque \bar{y}_w et \bar{y}_0 donnent des résultats très semblables dans ce cas. Pour les simulations dans la cellule 2 qui ne sont pas favorables à la pondération, la racine de l'erreur quadratique moyenne de l'estimateur

Tableau 3
Paramètres pour $[Y | M = m, X = c] \sim N(\beta_0 + \beta_1 c, \sigma^2)$

Association entre	β_1	σ^2	p^2
1. Forte	4,75	46	$\approx 0,80$
2. Moyenne	3,70	122	$\approx 0,48$
3. Faible	0,00	234	0,00

Nous avons simulé un biais d'échantillon répété de taille $n = 400$ et $n = 2\,000$ pour chaque combinaison de paramètres dans les tableaux 2 et 3. Les échantillons pour lesquels $n_{0c} = 0$ pour tout c ont été exclus, puisque l'estimation pondérée ne peut être calculée; en pratique, certaines cellules seraient probablement groupées dans de tels cas. Les nombres de simulations exclues sont présentés au tableau 4.

Tableau 4

Nombre de répliques exclues parce qu'une cellule ne contenait aucun répondant

Taux de réponse		70%	
Association	Association	entre Y et X	entre M et X
Forte	Forte	134	113
Moyenne	Faible	120	117
Moyenne	Faible	131	104
Moyenne	Faible	1	0

3.2 Comparaisons du biais, de la variance et de la racine de l'erreur quadratique moyenne, et leurs estimations

Les résultats sommaires du calcul empirique du biais et la racine de l'erreur quadratique moyenne (REQM) sont présentés au tableau 5. Nous pouvons comparer la valeur empirique de la REQM de la moyenne pondérée aux estimations suivantes, présentées au tableau 5, qui correspondent à la moyenne sur 1 000 répliques, à savoir la REQM estimée d'après l'équation (1) empirique de Kish, c'est-à-dire

$$\text{eqm}_{\text{Kish}}(\bar{y}_w) = (1 + L)s_y^2 / n_0, \quad \text{où } s_y^2 = \sum_{n_0}^i (y_i - \bar{y}_0)^2 / (n_0 - 1), \quad (15)$$

la REQM analytique provenant des équations (6) et (7), et la REQM estimée d'après les équations (11) et (13). Comme l'ont suggéré Oh et Scheuren (1983), nous incluons dans les deux dernières colonnes du tableau 5 le biais et la REQM empirique moyens d'une moyenne composite basée sur le choix, entre \bar{y}_w et \bar{y}_0 , de celle dont l'estimation de l'erreur quadratique moyenne fondée sur l'échantillon est la plus faible. Le biais empirique relativement au paramètre de population est présenté pour tous les estimateurs. Nous incluons aussi le biais et la RMSE de la moyenne avant suppression des cas de non-réponse.

$$B^2(\bar{y}_0) = \max\{0, (\bar{y}^w - \bar{y}_0)^2 - V_d\}$$

$$V_d = (n_1/n)^2 \left[\sum_{c=1}^C p_{1c}(\bar{y}_{0c} - \bar{y}_0)^2 / n_1 + \sum_{c=1}^C p_{0c}(\bar{y}_{0c} - \bar{y}_0)^2 / n_0 + \sum_{c=1}^C (p_{1c} - p_{0c})^2 / n_{0c} \right] + s^2 \sum_{c=1}^C (p_{1c} - p_{0c})^2 / n_{0c} \quad (12)$$

où $\bar{y}_{(1)}^w = \sum_{c=1}^C p_{1c} \bar{y}_{0c}$, et V_d estime la variance de $(\bar{y}^w - \bar{y}_0)$ et est inclus dans (12) comme correction du biais pour $(\bar{y}^w - \bar{y}_0)^2$ en tant qu'estimation de $B^2(\bar{y}_0)$, en suivant l'exemple de Little et coll. (1997). En outre

$$\text{eqm}(\bar{y}^w) = V(\bar{y}^w) = (1 + L)s^2/n_0 + \sum_{c=1}^C p_c(\bar{y}_{0c} - \bar{y}^w)^2/n. \quad (13)$$

Si nous soustrayons (11) de (13), l'écart entre les erreurs quadratiques moyennes de \bar{y}^w et \bar{y}_0 est alors estimé par

$$D = Ls^2/n_0 - (s_d^2 - s^2)/n_0 + \sum_{c=1}^C p_c(\bar{y}_{0c} - \bar{y}^w)^2/n - B^2(\bar{y}_0). \quad (14)$$

Il s'agit du perfectionnement de (1) que nous proposons, qui est représenté par le premier terme du deuxième membre de (14).

3. Étude en simulation

Nous incluons des simulations pour illustrer le biais et la

variance de la moyenne pondérée et non pondérée pour des ensembles de paramètres représentatifs chaque cellule du

tableau 1. Nous comparons aussi les approximations analytiques de l'erreur quadratique moyenne (MSE) dans les

Pourcentage de cas échantillonnés dans la cellule d'ajustement X et la cellule d'indication de données manquantes M

a. Taux de réponse global = 52 %

b. Taux de réponse global = 70 %												
Association		X	1	2	3	4	5	6	7	8	9	10
1.	Forte	M = 0	0,55	1,00	4,01	4,52	5,04	5,55	6,06	6,58	9,14	9,96
	Moyenne	M = 1	8,69	9,00	6,01	5,53	5,04	4,54	4,04	3,54	1,02	0,20
2.	Faible	M = 1	6,47	6,50	6,01	5,53	5,04	4,54	4,04	3,54	3,05	2,54
	Moyenne	M = 0	2,77	3,50	4,01	4,52	5,04	5,55	6,06	6,58	7,11	7,62
3.	Faible	M = 0	4,62	5,15	5,21	5,28	5,34	5,40	5,45	5,52	5,58	5,64
	Moyenne	M = 1	4,62	4,85	4,81	4,77	4,73	4,69	4,65	4,60	4,57	4,52
Association		X	1	2	3	4	5	6	7	8	9	10
entre M et X												

équations (6) et (7) et leurs estimations fondées sur un échantillon (11) et (13) à la MSE empirique sur des échantillons répétés.

3.1 Paramètres de superpopulation

Les spécifications de la simulation pour la loi conjointe de X et M sont décrites au tableau 2. L'échantillon suit approximativement une loi uniforme sur la variable de cellule d'ajustement X , qui compte $C = 10$ cellules. Nous choisissons deux taux de réponses marginaux, soit 70 %, qui correspond à une valeur de sondage typique et 52 %, qui est une valeur plus extrême pour accentuer les différences entre les méthodes. Nous simulons trois lois de M sachant X afin de modéliser une association forte, moyenne ou faible.

Les lois simulées de la variable d'intérêt Y sachant $M = m$, $X = c$ sont présentées au tableau 3. Elles ont toutes la forme

$$[Y | M = m, X = c] \sim N(\beta_0 + \beta_1 X, \sigma^2).$$

Trois ensembles de valeurs de (β_1, σ^2) sont simulés pour modéliser les associations forte, moyenne et faible entre Y et X . L'ordonnée à l'origine β_0 est choisie de sorte que la moyenne globale de Y soit $\mu = 26,3625$ pour chaque scénario.

Nous avons simulé un biais d'échantillon répété de taille $n = 400$ et $n = 2\,000$ pour chaque combinaison de paramètres dans les tableaux 2 et 3. Les échantillons pour lesquels $n_{0c} = 0$ pour tout c ont été exclus, puisque l'estimation pondérée ne peut être calculée; en pratique, certaines cellules seraient probablement groupées dans de tels cas. Les nombres de simulations exclues sont présentés au tableau 4.

f) L'équation (9) peut être appliquée au cas de la post-stratification sur les chiffres de population, en posant que n représente la taille de la population plutôt que la taille d'échantillon. Si l'on suppose que la population est grande, le deuxième terme de V_1 disparaît essentiellement, ce qui augmente la possibilité de réduction de la variance quand les variables formant les post-strates sont prédictives du résultat. Cette observation est une répétition de résultats antérieurs sur la post-stratification (Holt et Smith 1979; Little 1993).

Un sommaire qualitatif simple des résultats a) à f) de la section 2 est présenté au tableau 1, qui indique la direction du biais et la variance quand les associations entre les cellules d'ajustement et le résultat ainsi que l'indicateur de données manquants sont fortes ou faibles. De toute évidence, la pondération n'est efficace que pour les résultats qui sont associés à la variable de cellule d'ajustement, puisque autrement, elle accroît la variance sans réduction compensatoire du biais. Pour les résultats qui sont associés à la variable de cellule d'ajustement, la pondération accroît la précision et réduit également le biais si les variables de cellule d'ajustement sont reliées à la non-réponse.

Tableau 1

Effet de la pondération sur le biais et sur la variance d'une moyenne, selon la force de l'association des variables de cellule d'ajustement avec la non-réponse et le résultat			
Association avec le résultat			
Association avec la non-réponse	Faible	Forte	
	Cellule 1	Cellule 3	
Forte	Biais : ---	Biais : ---	Var : ---
	Cellule 2	Cellule 4	Var : ↑
Forte	Biais : ---	Biais : ---	Var : ↓
	Cellule 2	Cellule 4	Var : ↑

b) L'effet de la pondération sur la variance est représenté par $V_1 - V_2$.

c) Pour les résultats X qui ne sont pas reliés aux cellules d'ajustement, $\mu_{0c} = \mu_0$ pour tout c , $V_1 = 0$, et la pondération accroît la variance, puisque V_2 est positive. L'équation (10) réduit alors la version en population de la formule (1) de Kish. Les variables d'ajustement de cellule qui sont de bons prédicteurs de la non-réponse font plus de bien dans cette situation, puisqu'elles accroissent la variance des poids sans réduire le biais, mais il n'existe aucun compromis entre le biais et la variance pour ces résultats, puisqu'il n'y aucune réduction du biais.

d) Si la variable de cellule d'ajustement X n'est pas reliée à la non-réponse, alors λ est $O(1/n)$ et, par conséquent, V_2 a un ordre de variabilité plus faible que V_1 . Le terme V_1 a tendance à être positif, puisque $\sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2 \approx \sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2$, et le diviseur n dans le deuxième terme est plus grand que le diviseur m_0 dans le premier terme. Donc, dans ce cas, la pondération a tendance à n'avoir aucun effet sur le biais, mais elle réduit la variance dans la mesure où X est un bon prédicteur du résultat. Ceci contredit la notion selon laquelle la pondération accroît la variance. La « superefficacité » mentionnée plus haut qui résulte de l'estimation des poids de non-réponse à partir de l'échantillon est illustrée par le fait que, si les données sont manquantes complètement au hasard, alors le « vrai » poids de non-réponse est une constante pour toutes les unités répondantes. Par conséquent, la pondération au moyen des « vrais » poids produit (2), qui est moins efficace que la pondération par les poids « estimés », qui produit (3).

e) Si la variable de cellule d'ajustement est un bon prédicteur du résultat et également un prédicteur de la non-réponse, alors la valeur de V_2 est de nouveau faible, parce que la variance résiduelle σ^2 est réduite et celle de V_1 est généralement positive en vertu d'un argument semblable à celui exposé au point d). Le terme $\sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2$ peut s'écrire plus de $\sum_{c=1}^C \pi_{0c} (\mu_{0c} - \mu_0)^2$, parce que les poids sont moins susceptibles de déterminer le signe et la taille de V_1 . Donc, la pondération a tendance à la fois le biaiser et la variance dans ce cas-ci.

Il est utile de disposer d'estimations de la EQM de \bar{y}_0 et \bar{y}^w , qui peuvent être calculées d'après les données observées. Soit $s_{0c}^2 = \sum_{i \in c} (y_i - \bar{y}_{0c})^2 / (m_{0c} - 1)$ la variance d'échantillon des répondants dans la cellule c , $s^2 = \sum_{c=1}^C (m_{0c} - 1) s_{0c}^2 / (n_0 - C)$ la variance globale à l'intérieur des cellules et $s_0^2 = \sum_{i=1}^N (y_i - \bar{y}_0)^2 / (n_0 - 1)$ la variance d'échantillon totale des valeurs des réponses. Nous utilisons les expressions approximativement sans biais qui suivent, sous l'hypothèse que les données sont manquantes au hasard (MAR pour *missing at random*) :

$$\text{eqm}(\bar{y}_0) = B^2(\bar{y}_0) + V(\bar{y}_0), \quad (11)$$
$$\text{ou } V(\bar{y}_0) = s_0^2 / n_0 \text{ et}$$

cellules d'ajustement. Nous comparons le biais et l'erreur quadratique moyenne de (2) et (3) sous le modèle suivant, qui traduit les caractéristiques importantes du problème. Nous supposons que, sachant la taille d'échantillon n , les cas échantillonnés suivent une loi multinomiale sur le tableau de contingence $(C \times 2)$ basé sur la classification de M et X , avec les probabilités de cellule

$$\Pr(M=0, X=c) = \phi\pi_{0c}; \Pr(M=1, X=c) = (1-\phi)\pi_{1c},$$

où $\phi = \Pr(M=0)$ est la probabilité marginale de réponse.

La loi conditionnelle de X sachant $M=0$ et n_0 est multinomiale avec les probabilités de cellule $\Pr(X=c|M=0) = \pi_{0c}$, et la loi marginale de X sachant n est multinomiale

avec l'indice n et les probabilités de cellule

$$\Pr(X=c) = \phi\pi_{0c} + (1-\phi)\pi_{1c} = \pi_c,$$

disons. Nous supposons que la loi conditionnelle de Y sachant $M=m$, $X=c$ est de moyenne μ_{mc} et de variance constante σ^2 . Les moyennes de Y pour les répondants et les non-répondants sont

$$\mu_0 = \sum_{c=1}^C \pi_{0c} \mu_{0c}, \quad \mu_1 = \sum_{c=1}^C \pi_{1c} \mu_{1c},$$

respectivement, et la moyenne globale de Y est $\mu = \phi\mu_0 +$

$$(1-\phi)\mu_1.$$

Sous ce modèle, la moyenne et la variance conditionnelles

de \bar{y}_w sachant $\{p_c\}$ sont, respectivement, $\sum_{c=1}^C p_c \mu_{0c}$ et $\sigma^2 \sum_{c=1}^C p_c^2 / n_{0c}$. Donc, le biais de \bar{y}_w est

$$b(\bar{y}_w) = \sum_{c=1}^C \pi_c (\mu_{0c} - \mu_c),$$

où π_c et μ_c sont la proportion et la moyenne de population

de Y dans la cellule c . Cela peut s'écrire

$$b(\bar{y}_w) = \bar{\mu}_0 - \mu, \quad (4)$$

où $\bar{\mu}_0 = \sum_{c=1}^C \pi_c \mu_{0c}$ est la moyenne des répondants

« corrigée » pour les covariables et $\mu = \sum_{c=1}^C \pi_c \mu_c$ est la vraie moyenne de population de Y . La variance de \bar{y}_w est

égale à la somme de la valeur prévue de la variance conditionnelle et de la variance de son espérance conditionnelle, et est approximativement

$$V(\bar{y}_w) = (1 + \lambda) \sigma^2 \pi_c / n_0 + \sum_{c=1}^C \pi_c (\mu_{0c} - \bar{\mu}_0)^2 / n, \quad (5)$$

où $\lambda = \sum_{c=1}^C \pi_{0c} ((\pi_c / \pi_{0c} - 1)^2)$ est l'analogue de population de la variance des poids de non-réponse $\{w_c\}$, qui est identique à L dans l'équation (1), puisque les poids sont rééchantillonnés de sorte que leur moyenne soit égale à un. La formule de la variance pondérée dans Oh et Scheuren (1983), dérivée sous la perspective de quasi-randomisation, se réduit à (5) si l'on suppose que la variance dans les cellules est constante et que l'on ignore les corrections pour population finie et les termes d'ordre $1/n^2$. L'erreur quadratique moyenne (eqm) de \bar{y}_w est alors

$$\text{eqm}(\bar{y}_w) = b^2(\bar{y}_w) + V(\bar{y}_w). \quad (6)$$

L'erreur quadratique moyenne de la moyenne non pondérée (2) est donnée par

$$\text{eqm}(\bar{y}_0) = b^2(\bar{y}_0) + V(\bar{y}_0), \quad (7)$$

où

$$b(\bar{y}_0) = b(\bar{y}_w) + \mu_0 - \bar{\mu}_0, \quad (8)$$

est le biais et

$$V(\bar{y}_0) = \sigma^2 / n_0 + \sum_{c=1}^C \pi_{0c} (\mu_{0c} - \bar{\mu}_0)^2 / n, \quad (9)$$

est la variance. Donc, la différence (disons Δ) entre les

erreurs quadratiques moyennes est

$$\Delta = \text{eqm}(\bar{y}_0) - \text{eqm}(\bar{y}_w) = B + V_1 - V_2, \quad \text{où}$$

$$B = (\mu_0 - \bar{\mu}_0)^2 + 2(\mu_0 - \bar{\mu}_0)(\bar{\mu}_0 - \mu),$$

$$V_1 = \sum_{c=1}^C \pi_{0c} (\mu_{0c} - \bar{\mu}_0)^2 / n_0 - \sum_{c=1}^C \pi_c (\mu_{0c} - \bar{\mu}_0)^2 / n,$$

$$V_2 = \lambda \sigma^2 / n_0 \quad (10)$$

L'équation (10) et son interprétation détaillée fournissent les résultats principaux de l'article; il convient de souligner que les termes positifs dans (10) favorisent l'estimateur pondéré \bar{y}_w .

a) Le premier terme B représente l'effet sur l'erreur

quadratique moyenne de la réduction du biais due à l'ajustement sur les covariables. Il est d'ordre un et domine de plus en plus la EQM à mesure que augmente la taille de l'échantillon. Si $\mu \leq \bar{\mu}_0 < \mu_0$ ou $\mu_0 < \bar{\mu}_0 \leq \mu$, alors la pondération a réduit le biais de la moyenne des répondants et les deux composantes de B sont positives. En particulier, si les données manquant les sont au hasard (Rubin 1976; Little et Rubin 2002), en ce sens que les répondants constituent un échantillon aléatoire des cas échantillonnés dans chaque cellule c , alors $\bar{\mu}_0 = \mu$ et la pondération élimine le biais de la moyenne non pondérée. La correction du biais est

$$\mu_0 - \bar{\mu}_0 = \sum_{c=1}^C \pi_{0c} (1 - w_c) (\mu_{0c} - \mu_0),$$

si l'on ignore les différences entre les poids et leurs espérances. Il s'agit de zéro à $O(1)$ si la non-réponse n'est pas reliée aux cellules d'ajustement (auquel cas $w_c \approx 1$ pour tout c) ou que le résultat n'est pas relié aux cellules d'ajustement (auquel cas $\mu_{0c} \approx \mu_0$ pour tout c). Donc, une réduction importante du biais nécessite des variables d'ajustement reliées à la fois à la non-réponse et au résultat d'intérêt, fait qui a été souligné par plusieurs auteurs. On pense souvent que le conditionnement sur les caractéristiques observées des non-répondants réduira le biais, mais il

Kish (1992) présente une formule simple pour l'accroissement proportionnel de la variance dû à la pondération, disons L , sous l'hypothèse que la variance des observations est approximativement constante :

$$(1) \quad L = cv^2,$$

où cv est le coefficient de variation des poids des

répondants.

L'équation (1) est une bonne approximation si la variable de cellule d'ajustement est faiblement associée aux variables d'intérêt. Cependant, puisqu'elle donne une approximation de la variance plutôt que de l'erreur quadratique moyenne, elle ne mesure pas la réduction possible du biais dû à la non-réponse qui est l'objectif principal de la pondération et elle ne s'applique pas aux résultats qui sont associés à la

variable de cellule d'ajustement, pour lesquels la pondération pour la non-réponse peut en fait réduire la variance. Le fait que la pondération pour la non-réponse puisse réduire la variance est implicite dans la formule de Oh et Scheuren (1983) et est mentionné dans Little (1986) lorsque des cellules d'ajustement sont créées par stratification visant à prédire la moyenne. Ce fait se dégage aussi de la méthode connexe de post-stratification pour correction de la non-réponse (Holt et Smith 1979).

La variabilité des poids proprement dite ne se traduit pas

nécessairement par des estimations ayant une plus forte variance : une estimation pour laquelle la valeur de L est

élevée peut avoir une plus petite variance qu'une estimation dont la valeur de L est faible, comme l'illustrent les simulations présentées à la section 3. En outre, les situations où la pondération réduit le plus efficacement le biais dû à la non-réponse sont précisément celles où elle a tendance à réduire, et non à accroître, la variance et où l'équation (1) ne s'applique pas. Ces situations diffèrent du cas des poids de sondage et sont reliées à la « superefficacité » que l'on peut obtenir lorsque les poids sont estimés à partir de l'échantillon plutôt que d'être des constantes fixées; voir, par exemple, Robins, Rotnitzky et Zhao (1994).

Nous proposons un perfectionnement simple de l'équation (1), à savoir l'équation (14) données plus loin, qui reflète à la fois les composantes de biais et de variance, que la variable d'ajustement soit associée ou non aux résultats, et est, par conséquent, un indicateur plus précis de la valeur de la pondération des estimations et des variables de cellule d'ajustement. Dans les enquêtes polyvalentes comportant de nombreux résultats, l'approche type consiste à appliquer la même pondération pour la non-réponse à toutes les variables, en supposant implicitement que la valeur de la réduction du biais dû à la non-réponse pour certaines variables fait plus que compenser l'accroissement éventuel de la variance pour d'autres. Notre estimation empirique de l'erreur quadratique moyenne permet un

simple perfectionnement de cette stratégie, à savoir la réduction de la pondération au sous-ensemble de variables pour lesquelles elle réduit l'erreur quadratique moyenne estimée. Nous évaluons cette stratégie composite dans l'étude en simulation présentée à la section 3 et montrons qu'elle présente certains avantages par rapport à la pondération de toutes les variables de résultat. Comme nous le mentionnons à la section 4, d'autres approches présentent d'encore meilleures propriétés statistiques, mais elles produisent des poids différents pour chaque variable, ce qui rend leur mise en œuvre et leur explication aux utilisateurs des données d'enquête plus fastidieuses.

2. Répondération pour la non-réponse pour une moyenne

Supposons qu'on sélectionne un échantillon de n unités. Nous envisageons l'inférence pour la moyenne de population d'une variable étudiée X sujette à la non-réponse. Par souci de simplicité et pour nous concentrer sur la question de la correction pour la non-réponse, nous supposons que les unités sont sélectionnées par échantillonnage aléatoire simple. En général, les remarques faites ici au sujet de la pondération pour la non-réponse s'appliquent aussi à des plans de sondage complexes, quoique les détails techniques deviennent plus compliqués.

Nous supposons que les répondants et les non-répondants peuvent être classés dans C cellules d'ajustement d'après une covariable X . Soit M un indicateur de données manquantes dont la valeur est 0 pour les répondants et 1 pour les non-répondants. Soit n_{mc} le nombre d'individus échantillonnés avec $M = m$, $X = c$, $m = 0, 1$; $c = 1, \dots, C$. $n_{+c} = n_{0c} + n_{1c}$ représente le nombre d'individus échantillonnés dans la cellule c , $n_{0c} = \sum_{m=0}^1 n_{mc}$ et $n_{1c} = \sum_{m=1}^1 n_{mc}$, les nombres totaux de répondants et de non-répondants, et $p_c = n_{+c} / n$, $p_{0c} = n_{0c} / n_{0c}$, les proportions de cas échantillonnés et répondants dans la cellule c . Nous comparons deux estimations de la moyenne de population μ de X , à

savoir la moyenne non pondérée

$$(2) \quad \bar{y}_0 = \frac{1}{n} \sum_{c=1}^C p_{0c} \bar{y}_{0c},$$

où \bar{y}_{0c} est la moyenne pour les répondants dans la cellule c , et la moyenne pondérée

$$(3) \quad \bar{y}_w = \frac{1}{n} \sum_{c=1}^C p_c \bar{y}_{0c} = \frac{1}{n} \sum_{c=1}^C w_c p_{0c} \bar{y}_{0c},$$

où les répondants dans la cellule c sont pondérés par l'inverse du taux de réponse $w_c = p_c / p_{0c}$. L'estimateur par régression, où les valeurs manquantes sont imputées par la régression de X sur les indicateurs pour les

La pondération pour la non-réponse augmente-t-elle la variance des sondages?

Roderick J. Little et Sonya Vartivarian¹

Résumé

La pondération pour la non-réponse est une méthode courante de traitement de la non-réponse totale dans les sondages. Elle vise à réduire le biais dû à la non-réponse, mais produit souvent un accroissement de la variance. Par conséquent, son efficacité est souvent considérée comme un compromis entre le biais et la variance. Cette vision est cependant simpliste, car la pondération pour la non-réponse peut, en fait, réduire le biais ainsi que la variance. Pour réduire le biais de non-réponse, une covariable de pondération doit avoir deux caractéristiques : elle doit être corrélée à la probabilité de réponse, d'une part, et à la variable d'intérêt, d'autre part. Si cette deuxième caractéristique existe, la pondération peut réduire plutôt qu'augmenter la variance d'échantillonnage. Nous présentons une analyse détaillée du biais et de la variance dans le cas d'une pondération pour l'estimation d'une moyenne de sondage au moyen de cellules d'ajustement. L'analyse donne à des variables d'intérêt, la prédiction de la propension à répondre est un objectif secondaire, quoiqu'utile. Nous proposons des estimations empiriques de la racine carrée de l'erreur quadratique moyenne pour déterminer dans quelles circonstances la pondération est efficace et nous les évaluons au moyen d'une étude en simulation. Un estimateur composite simple fondé sur la racine de l'erreur quadratique moyenne empirique donne de meilleurs résultats que l'estimateur pondéré dans les simulations.

Mots clés : Données manquantes; correction pour la non-réponse; poids de sondage; non-réponse à une enquête.

1. Introduction

Dans la plupart des enquêtes, certaines personnes échantillonnées ne fournissent pas d'information parce qu'on n'a pas pu prendre contact avec elles ou qu'elles ont refusé de répondre (non-réponse totale). La méthode la plus courante de correction de la non-réponse totale est la pondération, où les répondants et les non-répondants sont classés dans des cellules d'ajustement d'après des données sur des covariables dont les valeurs sont connues pour toutes les unités échantillonnées, et un poids de non-réponse est calculé pour les cas compris dans une cellule proportionnellement à l'inverse du taux de réponse dans la cellule. Souvent, on multiplie le poids de sondage par ces poids de non-réponse et on normalise le poids global de sorte que la somme des poids des cellules soit égale au nombre de répondants dans l'échantillon. Oh et Scheuren (1983) donnent une bonne vue d'ensemble de la pondération pour la non-réponse. Une approche apparemment est la post-stratification (Holt et Smith 1979), qui s'applique lorsque la distribution de la population entre les cellules d'ajustement peut être déterminée d'après des sources externes, comme le recensement. Le poids est alors proportionnel au ratio du chiffre de population au nombre de répondants dans une cellule.

La pondération pour la non-réponse, ou pondération, est considérée principalement comme un moyen de réduire le biais dû à la non-réponse totale. Ce rôle de la

pondération est analogue à celui des poids de sondage et est relié à la propriété d'absence de biais par rapport au plan de sondage de l'estimateur du total d'Horvitz-Thompson (Horvitz et Thompson 1952), où les unités sont pondérées par l'inverse de leur probabilité de sélection. La pondération pour la non-réponse peut être considérée comme une extension naturelle de cette idée, où les unités comprises dans l'échantillon sont pondérées par l'inverse de leur probabilité d'inclusion, estimée comme étant le produit de la probabilité de sélection et de la probabilité de réponse, sachant que l'unité a été sélectionnée; l'inverse de la seconde probabilité est le poids de non-réponse. Bien que certains modélisateurs soutiennent que la pondération en vue de corriger le biais n'est pas nécessaire dans les modèles où la pondération n'est pas associée aux variables d'intérêt, en pratique, peu d'entre eux sont prêts à émettre une hypothèse aussi forte.

Les poids de sondage réduisent le biais au prix d'un accroissement de la variance, si la variance du résultat est constante. Étant donné l'analogie entre les poids de non-réponse et les poids de sondage, il paraît plausible que la pondération pour la non-réponse réduise aussi le biais au prix d'un accroissement de la variance des estimations par sondage. La notion de compromis entre le biais et la variance est abordée dans certaines discussions de la pondération pour la non-réponse (Kallon et Kasprzyk 1986; Kish 1992; Little, Lewitzky, Heeringa, Lepkowski et Kessler 1997).

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Fay, R.E. (1992). When are imputations from multiple imputation valid. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquêtes manquantes. *Techniques d'enquête*, 12, 1-17.

Lee, H., Rancourt, E. et Särndal, C.-E. (1995). Jackknife variance estimation for data with imputed values. *Proceedings of the Statistical Society of Canada Survey Methods Section*, 111-115.

Lee, H., Rancourt, E. et Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. Dans *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), Chapitre 21, New York: John Wiley & Sons Inc.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons Inc.

Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input. (avec discussion). *Statistical Science*, 9, 538-573.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1996). Multiple imputation after 18+ years (avec discussion). *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., et Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with nonignorable nonresponse. *Journal of the American Statistical Association*, 81, 361-374.

Rust, K., et Rao, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medicine*, 5, 381-397.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite estimation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Brick, Jones, Kalton et Valliant : Estimation de la variance avec imputation hot deck

Rockville, MD: Westat, Inc.

RN95127001 to the National Center for Education Statistics.

study of three methods of variance estimation with hot deck imputation for stratified samples. Préparé sous contrat No.

Brick, J.M., Jones, M., Kalton, G. et Valliant, R. (2004). A simulation study of three methods of variance estimation with hot deck imputation for stratified samples. Préparé sous contrat No.

Brick, J.M., Kalton, G. et Kim, J.K. (2004). Estimation de variance pour l'imputation hot deck à l'aide d'un modèle. *Techniques d'enquête*, 30, 63-72.

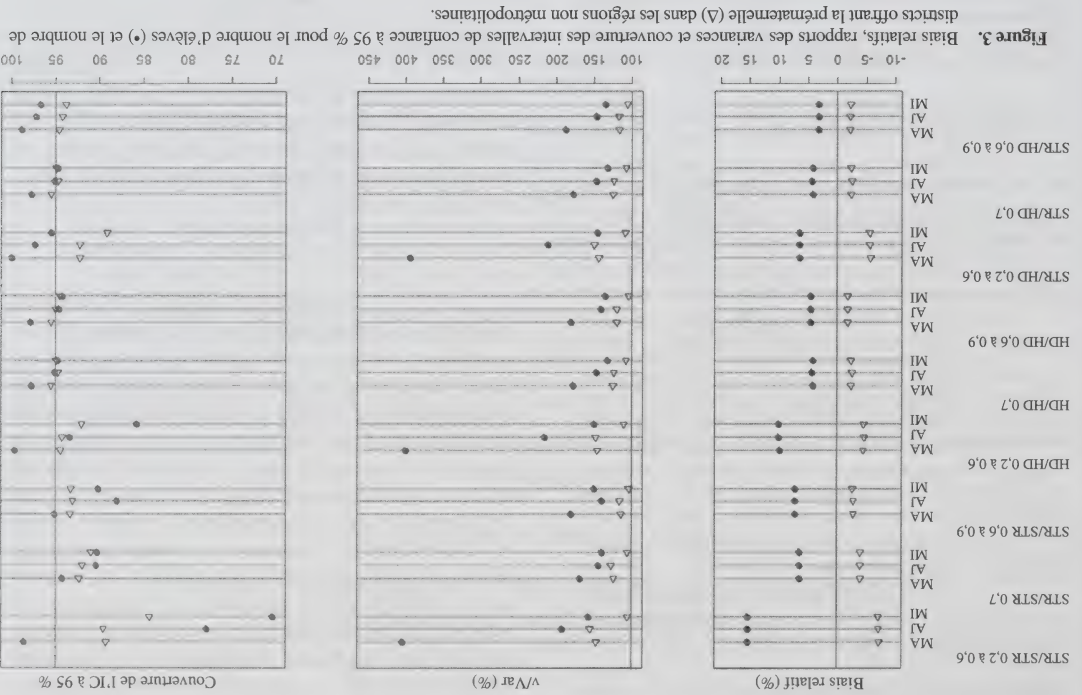
Les auteurs remercient l'Institut for Education Sciences du National Center for Education Statistics d'avoir appuyé cette étude, en particulier Marilyn Seastrom. Nous tenons aussi à remercier les examinateurs de leurs commentaires constructifs.

nos simulations prouvent son importance.

d'imputation sont susceptibles de présenter un biais important. Si les estimations ponctuelles sont fortement biaisées, les méthodes peuvent produire des intervalles de confiance dont la couverture est de loin inférieure au taux nominal. Les analyses des ensembles de données imputées devraient déterminer si la méthode d'imputation qui a été utilisée produira vraisemblablement des estimations appproximativement sans biais, particulièrement pour les estimations par domaine. Sinon, il pourrait être nécessaire qu'ils réimputent les réponses manquantes pour obtenir des estimations ponctuelles moins biaisées. Recommander aux imputeurs de tirer parti d'autant de variables explicatives que possible dans le processus d'imputation n'est pas un conseil nouveau, mais les conclusions qui se dégagent de

Remerciements

Bibliographie



5. Conclusion

Nos simulations avaient pour but d'examiner les propriétés de trois estimateurs de la variance des estimations de totaux imputés d'après un plan d'échantillonnage stratifié sous divers mécanismes de réponse avec imputation hot deck pondérée. Les conditions des simulations reflètent celles auxquelles on peut s'attendre en pratique en ce sens que les hypothèses qui sous-tendent les méthodes sont violées de diverses façons. Les trois méthodes produisent des estimations nettement meilleures que l'estimateur naïf de la variance. Elles donnent toutes trois de très bons résultats quand les estimations ponctuelles sont sans biais. Si le biais dans les estimations ponctuelles est important, aucune des méthodes ne produit des intervalles de confiance dont la couverture correspond au taux nominal. L'obtention de taux de couverture médiocres pour des estimations ponctuelles biaisées n'est pas inattendue, puisqu'il en est également ainsi quand il n'y a pas de données manquantes. Lorsque le biais des estimations ponctuelles est relativement faible, les taux de couverture réels obtenus pour les trois méthodes d'estimation de la variance sont parfois supérieurs et parfois inférieurs au taux nominal. Dans ce cas, la tendance des trois méthodes à surestimer la variance produit souvent des taux de couverture proches du taux nominal.

Les résultats de la présente étude donnent aux praticiens de l'imputation hot deck des preuves empiriques que toutes les méthodes d'estimation de la variance donnent de bons résultats en cas d'échantillonnage à un seul degré à condition que l'estimation ponctuelle soit sans biais, même si d'autres hypothèses sont violées. Les estimations pour des domaines qui ne sont pas pris en compte dans le scénario

Les faibles taux de réponse sont associés à un taux de couverture trop faible, dû en grande partie aux biais plus importants dans les estimations ponctuelles.

En général, les écarts entre les taux de couverture obtenus pour les trois méthodes sont faibles et ne permettent pas d'affirmer que l'une des méthodes est supérieure aux autres en général. Pour des taux de réponse très faibles, la longueur moyenne des intervalles de confiance pour la méthode MI est appréciablement plus grande que celle observée pour les méthodes MA et AJ, mais l'utilisation d'un plus grand nombre d'ensembles d'imputations avec la méthode MI pourrait corriger ce problème. Il convient toutefois de souligner que ces simulations ne portent que sur le cas de l'échantillonnage à un degré. Il pourrait exister des écarts entre les longueurs des intervalles de confiance produits par les diverses méthodes en cas d'échantillons en grappes. Cette possibilité devrait faire l'objet d'études futures.

0,2 et 0,6, dont le taux de couverture pour le nombre d'élèves peut être aussi faible que 86 %.

Pour les scénarios STR/STR, la figure 2 indique que toutes les méthodes ont tendance à produire une couverture inférieure au taux nominal pour le nombre d'élèves et supérieures au taux nominal pour le nombre de districts offrant la prématernelle. L'écart entre les taux de couverture pour les deux variables est dû à la taille du biais relatif des estimations ponctuelles et des estimations de la variance.

Si nous passons aux estimations pour le domaine NMSA à la figure 3, il convient de souligner que la situation de région métropolitaine n'est explicitement incluse ni dans la définition de STR ni dans celle de HD, bien qu'elle soit clairement corrélée à la taille et, donc, à STR. Pour tous les scénarios, les estimations ponctuelles du nombre d'élèves dans le domaine NMSA présentent un biais positif important. La couverture des intervalles de confiance MA est uniformément égale ou supérieure au taux nominal,

principalement à cause du biais positif extrême dans les estimations de la variance. La couverture des intervalles AJ s'approche du taux nominal pour les scénarios HD/HD et STR/HD, mais est inférieure à ce taux pour les trois scénarios STR/STR. Les profils de couverture pour les intervalles de confiance MI sont semblables à ceux observés pour la méthode AJ, sauf que la couverture des intervalles MI est appréciablement inférieure au taux nominal pour le scénario HD/HD avec taux de réponse de 0,2 à 0,6.

Les estimations ponctuelles du nombre de districts avec prématernelle dans le domaine NMSA ont un biais relatif négatif modéré pour chacun des neuf scénarios. Pour les trois méthodes d'estimation de la variance, la couverture des intervalles de confiance s'approche du taux nominal, sans la surcouverture observée pour les estimations pour le domaine NE.

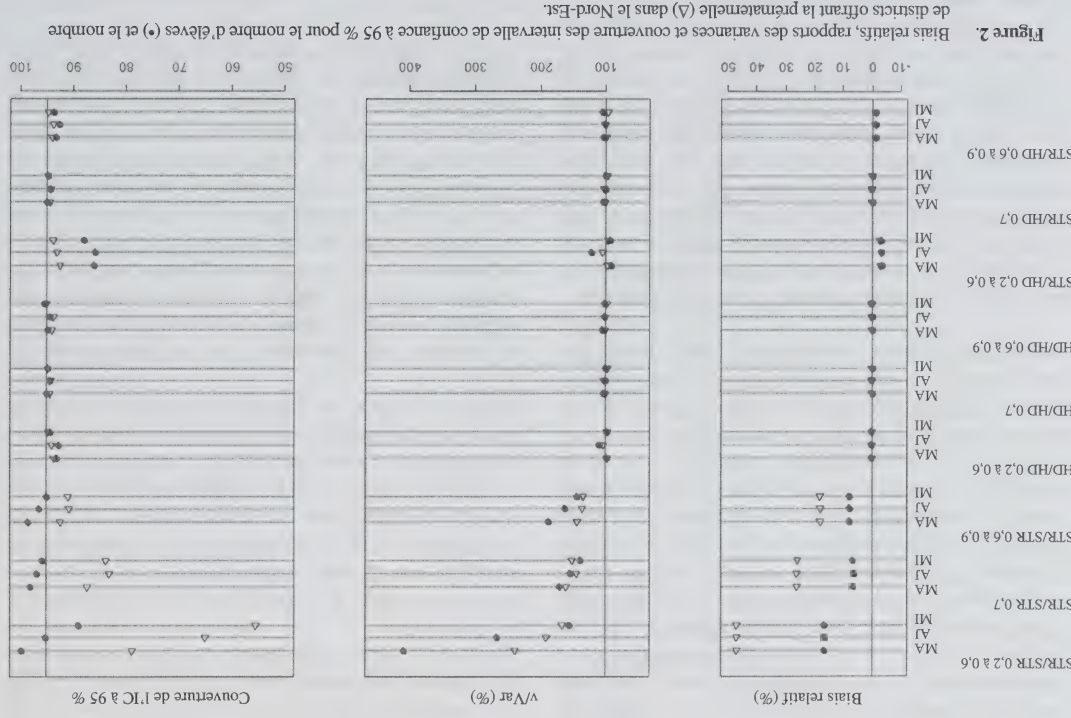


Figure 2. Biais relatifs, rapports des variances et couverture des intervalles de confiance à 95 % pour le nombre d'élèves (*) et le nombre de districts offrant la prématernelle (Δ) dans le Nord-Est.

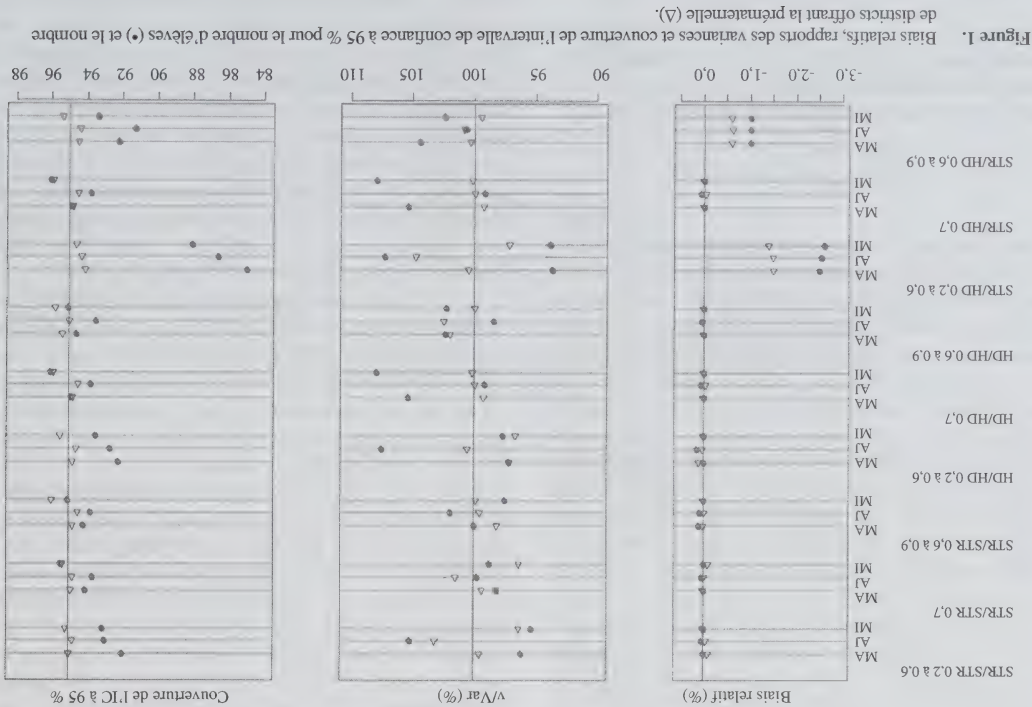
4. Résultats des simulations

À la présente section, nous présentons les principaux résultats des simulations, en commençant par la perfor-
mance des trois méthodes d'estimation de la variance des
estimations pour l'ensemble de la population, puis nous
donnons les résultats pour les estimations par domaine. Les
résultats essentiels sont résumés graphiquement ici, mais les
tableaux contenant les données détaillées peuvent être
consultés dans Brick, Jones, Kalton et Valliant (2004).

4.1 Estimations pour l'ensemble de la population

La figure 1 illustre les résultats des simulations pour
l'estimation du nombre total d'élèves et du nombre de
districts offrant la prématernelle d'après les 10 000 échan-
tillons pour chacun des 9 scénarios de simulation. \bar{Y} sont
présentés le biais relatif de l'estimateur imputé, l'estimation
moyenne de la variance en pourcentage de la variance empi-
rique et le taux de couverture de l'intervalle de confiance.

Le graphique des ratios de l'estimation moyenne de la
variance à la variance empirique (v/Var dans les figures)
pour les trois méthodes montre que ces estimations
présentent un biais assez faible dans la plupart des cas,
compris dans la fourchette de plus ou moins 8 % par rapport
à la variance réelle simulée. Bien que les ratios pour toutes
les méthodes varient entre les neuf scénarios, les ratios MI
sont un peu plus variables que ceux calculés pour les deux
autres méthodes.



situation pour l'estimation ponctuelle et l'estimation de la variance est la même que celle mentionnée plus haut pour les totaux globaux de population.

De façon générale, nous avons construit les scénarios de simulation de sorte que les cellules hot deck n'englobent pas les domaines, afin de refléter le fait qu'en pratique, il est essentiellement impossible d'intégrer tous les domaines dans un schéma d'imputation. Plus précisément, dans les simulations, les districts de la région du Nord-Est (NE) et ceux des régions statistiques non métropolitaines (NMSA) ne sont pas reliés aux définitions de strate du tableau 1 (qui sont utilisées comme cellules hot deck dans certains cas). En outre, les districts compris dans le domaine NMSA peuvent se retrouver dans toutes les cellules HD. Cependant, le domaine NE est un sous-ensemble de quatre cellules HD. Donc, la définition des cellules HD concorde plus avec l'estimation des totaux de domaine NMSA.

3.4 Statistiques sommaires

Le biais relatif d'une estimation ponctuelle est estimé par $rel(biais(\theta_j) = biais(\theta_j) / \theta_N$, où $biais(\theta_j) = \sum_s (\hat{\theta}_{js} - \theta_N) / 10\,000$, $\hat{\theta}_{js}$ est l'estimation d'échantillon s , et θ_N est le paramètre de population finie. La variance empirique de $\hat{\theta}_j$ est $Var(\hat{\theta}_j) = \sum_s (\hat{\theta}_{js} - \hat{\theta}_j)^2 / 10\,000$, où $\hat{\theta}_j = \sum_s \hat{\theta}_{js} / 10\,000$. L'estimation moyenne de la variance pour une méthode particulière est $v = \sum_s v_s / 10\,000$, où v_s est la variance estimée pour l'exécution de simulation s .

Les pourcentages d'intervalles qui comprennent θ_N sont fondés sur les intervalles de confiance à 95 % nominaux $(\theta_j \pm t_{1/2}^{1/2})$ calculés pour chacune des 10 000 simulations pour chaque scénario de simulation. Un élément dont il faut tenir compte ici est la précision des estimations de la variance d'après un plan d'échantillonnage stratifié non proportionnel et son incidence sur la décision d'utiliser une approximation normale ou bien des intervalles t pour autres combinaisons STR/HD. La théorie de la méthode AJ d'estimation de la variance de population a été élaborée en supposant seulement que le modèle de probabilité de réponse est vérifié. Les théories MA et MI reposent sur l'hypothèse que les deux modèles sont vérifiés.

La possibilité d'utiliser uniquement le modèle de probabilité de réponse et l'imputation hot deck pondérée pour produire des estimations sans biais des totaux de population ne s'étend généralement pas à l'estimation des totaux de domaine. Si le domaine recoupe les cellules hot deck, il est nécessaire d'utiliser un modèle de population qui suppose que l'espérance des valeurs de domaine est la même que celle des valeurs hors domaine dans chaque cellule hot deck. Cependant, si les cellules hot deck sont définies de façon que chaque domaine comprenne la population complète dans un sous-ensemble des cellules hot deck, alors la

Trois des quatre combinaisons possibles de mécanisme de réponse (cellules STR ou HD) et de formation de cellules hot deck (cellules STR ou HD) ont été étudiées dans les simulations. Nous nommons ces combinaisons STR/STR, HD/HD et STR/HD, où le premier ensemble de lettres désigne le mécanisme de réponse et le deuxième, le type de cellule hot deck. Les trois ensembles de valeurs du taux de réponses étaient de 0,2 à 0,6 espacées uniformément entre les cellules de réponse, une valeur constante de 0,7 dans toutes les cellules. Les trois combinaisons cellule de réponse/cellule hot deck et les trois ensembles de taux de réponse ont généré neuf scénarios de simulation distincts pour chaque estimation.

3.3 Hypothèses concernant les modèles de réponse et la structure de la population

Deux modèles interviennent dans les simulations. Le modèle de population repose sur l'hypothèse que les valeurs de y dans chaque cellule hot deck sont indépendantes et ont la même espérance. Le modèle de réponse repose sur l'hypothèse qu'il existe une probabilité de réponse uniforme dans chaque cellule hot deck. Si les deux modèles tiennent, alors l'utilisation d'une imputation hot deck non pondérée produira une estimation sans biais du total global de population. Par contre, si l'on suppose que seul le modèle de réponse est vérifié, alors l'utilisation d'une imputation hot deck pondérée est nécessaire pour produire une estimation sans biais de ce total. Puisque l'imputation hot deck pondérée est utilisée dans les simulations, il suffit que le modèle de probabilité de réponse soit satisfait pour obtenir une estimation ponctuelle sans biais du total global de population. Le modèle de probabilité de réponse est vérifié pour toutes les combinaisons STR/STR et HD/HD, ainsi que pour la combinaison STR/HD avec taux de réponse constant; cependant, il ne tient pas pour les deux autres combinaisons STR/HD. La théorie de la méthode AJ d'estimation de la variance de population a été élaborée en supposant seulement que le modèle de probabilité de réponse est vérifié. Les théories MA et MI reposent sur l'hypothèse que les deux modèles sont vérifiés.

La possibilité d'utiliser uniquement le modèle de probabilité de réponse et l'imputation hot deck pondérée pour produire des estimations sans biais des totaux de population ne s'étend généralement pas à l'estimation des totaux de domaine. Si le domaine recoupe les cellules hot deck, il est nécessaire d'utiliser un modèle de population qui suppose que l'espérance des valeurs de domaine est la même que celle des valeurs hors domaine dans chaque cellule hot deck. Cependant, si les cellules hot deck sont définies de façon que chaque domaine comprenne la population complète dans un sous-ensemble des cellules hot deck, alors la

Puisque l'utilisation de 1,96 avec la méthode MI a donné des intervalles dont la couverture est nettement trop faible, nous utilisons la loi t avec λ degrés de liberté pour les intervalles de confiance MI.

$$\lambda = (M - 1) \left(1 + \frac{M}{M + 1} \frac{B}{U} \right)^2.$$

la base de sondage. Pour générer des valeurs manquantes, nous avons attribué des indicateurs de réponse aux unités échantillonnées comprises dans les « cellules de réponse ». Dans certains cas, ces dernières sont les strates d'échantillonnage, et sont nommées cellules STR, tandis que dans d'autres, il s'agit de cellules nommées cellules HD. Ces dernières ont été définies par croisement de quatre régions géographiques et d'une catégorisation à quatre niveaux du nombre d'enseignants équivalents temps plein dans le district. Les cellules HD sont plus ou moins corrélées aux strates d'échantillonnage, mais chacune contient des unités provenant de plus d'une strate.

Dans une cellule de réponse donnée, nous avons réparti aléatoirement les unités échantillonnées entre les catégories chaque type de cellule de réponse, nous avons choisi trois scénarios pour attribuer les taux de réponses manquantes. Dans deux scénarios, ce taux variait selon la cellule de réponse, tandis que dans le troisième, il était constant dans toutes les cellules.

Nous avons réalisé les simulations en tirant d'abord un échantillon aléatoire simple stratifié d'après les tailles d'échantillon de strate du tableau 1. Après avoir sélectionné l'échantillon, nous avons attribué aléatoirement la situation de réponse (répondant/non-répondant) à chaque unité échantillonnée conformément au scénario de réponse donné. Pour les méthodes MA et AJ, nous avons utilisé les procédures d'imputation hot deck pondérée décrites plus haut pour imputer les valeurs manquantes. Pour la méthode MI, nous avons d'abord créé un groupe de donneurs en utilisant l'ABB pondéré, puis nous avons utilisé la méthode hot deck pondérée pour chacun des $M = 5$ ensembles de données imputés. Nous avons calculé les nombres totaux estimés d'élèves et de districts avec prématurée pour l'échantillon simulé avec valeur imputée, ainsi que les

Tableau 1 Définitions des strates, chiffres de population, tailles d'échantillon, taux d'échantillonnage, moyennes et écarts-types du nombre d'élèves et proportions de districts offrant une prématurée.

Strate	Taille du district	Situation de pauvreté	N_h	n_h	Taux d'échant.	Moyenne	Écart-type	Proportion avec prématurée
1	1	1	615	32	0,0520	270,0	155,0	0,44
2	1	2	1 147	59	0,0514	263,3	175,0	0,49
3	1	3	1 292	66	0,0511	243,5	142,5	0,49
4	2	1	1 720	111	0,0645	1 607,2	837,0	0,44
5	2	2	2 305	149	0,0646	1 429,7	784,1	0,52
6	2	3	1 893	122	0,0644	1 427,8	788,8	0,63
7	3	1	692	75	0,1084	4 695,3	1 360,6	0,35
8	3	2	579	63	0,1088	4 728,5	1 365,0	0,51
9	3	3	527	57	0,1082	4 591,8	1 380,3	0,63
10	4	1	342	83	0,2427	16 003,4	12 670,2	0,51
11	4	2	449	110	0,2450	17 577,3	14 246,7	0,58
12	4	3	380	93	0,2447	19 331,8	16 142,7	0,68
Total			11 941	1 020		3 237,9	6 770,5	0,52

Une caractéristique de la conception des simulations est que les moyennes des deux domaines considérés diffèrent souvent beaucoup des moyennes de population complète selon la strate et la cellule HD. Une remarque importante en ce qui concerne les estimations par domaine est que les imputations ont été faites en sélectionnant des donneurs à partir de l'ensemble des répondants dans une cellule hot deck, sans reconnaître précisément le domaine comme nous l'avons estimé le total pour un domaine par $\theta_i = \sum_{j \in A_{ij}} \delta_{ij} w_{ij}^* + \sum_{j \in A_{ij}^*} \delta_{ij} w_{ij}^*$ où $\delta_i = 1$ si l'unité i est dans le domaine et 0 autrement.

les simulations, nous avons modifié l'ABB afin de tenir compte de l'échantillonnage ppsar des enregistrements donneurs. Nous avons créé un groupe d'enregistrements donneurs pour l'ABB dans chaque cellule en sélectionnant les répondants avec probabilité proportionnelle à w_i . (Aucun article publié ne discute de l'application des méthodes ABB avec des poids inégaux. A posteriori, nous pensons qu'une méthode ABB non pondérée aurait été préférable. L'utilisation d'un ABB non pondéré avec une imputation hot deck ppsar produit des estimations ponctuelles sans biais des totaux de population sous le modèle de probabilité de réponse).

3. Conception de l'étude par simulation

3.1 Description de la population étudiée et du plan d'échantillonnage

La base de sondage pour les simulations est un sous-ensemble du fichier des districts scolaires publics extraits du Common Core of Data (CCD) de 1999-2000 assemblé par le U.S. National Center for Education Statistics. La base de sondage finale comprend 11 941 districts.

Pour les simulations, nous avons sélectionné un échantillon de 1 020 districts scolaires conformément à un plan d'échantillonnage aléatoire simple stratifié. Nous avons créé 12 strates par croisement de 4 catégories de nombre d'élèves (taille du district) et 3 catégories de pourcentage d'élèves se trouvant au seuil de pauvreté ou sous celui-ci (situation de pauvreté). Les strates et les nombres de districts dans la base de sondage sont présentés au tableau 1. Celui-ci donne aussi les tailles d'échantillon de strate et les taux d'échantillonnage utilisés dans les simulations.

Le tableau contient également les moyennes et les écarts-types de strate pour les deux variables étudiées, c'est-à-dire le nombre d'élèves dans le district et le nombre de district dont le niveau d'enseignement le plus faible est la prématernelle. Nous avons choisi d'étudier ces variables, parce qu'elles sont typiques de nombreuses estimations calculées d'après le genre de plan de sondage susmentionné. En plus des estimations pour la population dans son ensemble, nous avons estimé la valeur de deux variables étudiées pour deux domaines, définis comme étant les districts situés dans la région du Nord-Est et ceux situés dans des régions non métropolitaines. Les moyennes pour ces domaines sont très différentes des moyennes pour la population dans son ensemble pour les deux variables étudiées.

3.2 Mécanismes de génération des données manquantes et méthodes d'imputation

Par construction, l'information sur les deux variables étudiées est disponible pour tous les districts compris dans

échantillons sont affectées à l'un des groupes. Alors, l'estimateur de la variance par le jackknife ajusté groupé est

$$\hat{V}_{AJ} = \sum_{h=1}^H \sum_{k=1}^K n_{h(k)}^2 (\hat{\theta}_{h(k)}^{(k)} - \hat{\theta}_I)^2,$$

où $n_{h(k)}$ est le nombre d'unités d'échantillonnage dans la strate de variance combinée $h^*, n_{h(k)}$ est le nombre d'unités retenues dans la strate h^* quand les unités du groupe k sont supprimées et, correspondant à $\hat{\theta}_{h(k)}^{(k)}, \hat{\theta}_{h(k)}^{(k)}$ est l'estimation imputée ajustée pour l'ensemble de la population quand les unités du groupe k dans la strate h^* sont supprimées. Les unités retenues provenant de la strate de plan de sondage h qui figurent dans la strate de variance combinée h^* reçoivent un poids de rééchantillonnage $w_{h(k)}^{(k)} = n_{h(k)}^2 / n_{h(k)}$.

La méthode AJ repose sur l'hypothèse d'un modèle à probabilité de réponse uniforme dans chaque cellule hot deck, mais, contrairement à la méthode MA, ne nécessite pas d'hypothèse concernant la loi de distribution. Sous le modèle de probabilité de réponse uniforme sans hypothèse concernant la loi de distribution, l'utilisation d'une méthode hot deck pondérée est nécessaire pour produire des estimations imputées sans biais.

Quand ils ont élaboré la théorie de la méthode AJ, Rao et Shao (1992) ont supposé que les facteurs f_{cfd} étaient ignorables. Cependant, dans les simulations, ils ne le sont pas dans certaines strates, leur valeur variant d'environ 0,05 à 0,24. Shao et Steel (1999), ainsi que Lee, Rancourt et Sæmdal (1995) donnent des méthodes pour tenir compte des f_{cfd} non négligeables. Dans la simulation, nous avons utilisé l'ajustement pour les f_{cfd} proposé par Lee, Rancourt et Sæmdal (1995) parce qu'il est facile à appliquer. Sans ajustement pour les f_{cfd} , l'estimateur de la variance AJ surestime considérablement les variances dans les simulations.

2.3 Imputation multiple

L'imputation multiple (MI) est décrite en détail dans Rubin (1987), ainsi que dans Little et Rubin (2002). Le résumé présenté ici a trait à son application en présence d'imputation hot deck. Comme pour l'approche assistée par modèle, nous supposons que, dans les cellules hot deck, les réponses manquent de façon aléatoire et que les y sont des variables aléatoires indépendantes de moyenne et variance communes. Pour chaque unité pour laquelle une valeur manque, M valeurs sont imputées, ce qui crée M ensembles de données complets. Pour éviter de sous-estimer les variances par la méthode MI, il faut modifier la méthode hot deck. Rubin et Schenker (1986) ont proposé le bootstrap bayésien approximatif (ABB) pour l'échantillonnage aléatoire simple avec imputation hot deck en cas d'utilisation de la méthode MI. Pour

2.1 Estimation de la variance assistée par modèle

L'approche assistée par modèle (MA) avec imputation hot deck repose sur l'hypothèse que les données manquent aléatoirement dans les cellules et qu'un modèle pour la génération des y est vérifié. Un modèle naturel en cas d'imputation hot deck est que les y_i sont générés de façon indépendante et identique dans les cellules hot deck, c'est-à-dire, $y_i^{ht} \sim (\mu_g, \sigma_g^2)$ pour la cellule g . Sous l'approche assistée par modèle, les inférences dépendent de la validité des hypothèses du modèle.

Sandall (1992) a décomposé la variance totale de l'estimateur imputé en trois composantes notées V_{SAM}^{imp} , V_{MIX}^{imp} et V_{MIX}^{MA} . Les estimateurs utilisés pour ces composantes dans les simulations sont ceux donnés dans Brick, Kalton et Kim (2004). L'estimateur MA de la variance est égal à la somme des estimations composantes : $V_{MA}^{imp} = V_{SAM}^{imp} + V_{MIX}^{imp} + 2V_{MIX}^{MA}$. Les estimateurs V_{MIX}^{imp} et V_{MIX}^{MA} requièrent un estimateur de la variance élémentaire dans chaque cellule hot deck. Puisque les simulations ont indiqué que la différence entre les estimateurs pondérés et non pondérés était faible, nous ne discutons que de l'estimateur pondéré de σ_g^2 , c'est-à-dire $\hat{\sigma}_g^2 = n_{Rg}^{-1} (n_{Rg} - 1)^{-1} \sum_{i \in A_{Rg}} w_i^{-1} (y_i - \bar{y}_{Rg})^2 \times (\sum_{i \in A_{Rg}} w_i^{-1})^{-1}$, avec $\bar{y}_{Rg} = \sum_{i \in A_{Rg}} w_i^{-1} y_i / (\sum_{i \in A_{Rg}} w_i^{-1})$.

2.2 Estimation de la variance par le jackknife ajusté

L'estimateur de la variance par le jackknife ajusté (AJ) de Rao et Shao (1992) pour un échantillon stratifié avec imputation et facteurs de correction pour population finie ($fcpf$) ignorable est

$$\hat{V}(\hat{\theta}_I) = \sum_{h=1}^H \sum_{n_h=1}^n \frac{1}{n_h} (\hat{\theta}_{(k)}^{h_{(k)}} - \hat{\theta}_I)^2,$$

où n_h est le nombre d'unités échantillonnées dans la strate h ,

$$\hat{\theta}_{(k)}^{h_{(k)}} = \sum_{g=1}^G \left\{ \sum_{(ht) \in A_{Rg}} w_{ht}^{(k)} y_{ht} + \sum_{(ht) \in A_{Mg}} w_{ht}^{(k)} (y_{ht}^* + \hat{y}_{Rg}^{(k)} - \bar{y}_{Rg}) \right\}$$

est l'estimateur ajusté quand l'unité k est omise,

$$\hat{y}_{Rg}^{(k)} = \sum_{(ht) \in A_{Rg}} w_{ht}^{(k)} y_{ht} / \sum_{(ht) \in A_{Rg}} w_{ht}^{(k)}, \quad \bar{y}_{Rg}^{(k)} = \sum_{(ht) \in A_{Rg}} w_{ht}^{(k)} y_{ht} / \sum_{(ht) \in A_{Rg}} w_{ht}^{(k)},$$

$w_{ht}^{(k)}$ est le poids de l'unité ht ajusté pour tenir compte de l'omission de l'unité k . La notation $(ht) \in B$ dénote que l'unité i dans la strate h fait partie de l'ensemble B . Cette procédure requiert le calcul de $\sum n_h$ estimations répétées, $\hat{\theta}_{(k)}^{h_{(k)}}$. Une stratégie utilisée fréquemment pour réduire les calculs consiste à combiner les unités en strates de variance commune combinée et k un groupe d'unités d'échantillonage dans la strate combinée. Toutes les unités

qui sous-tendent les méthodes d'estimation de la variance que nous étudions reposent chacune sur l'hypothèse que les données manquent au hasard dans chaque cellule hot deck. En outre, la méthode assistée par modèle (MA) et la méthode d'imputation multiple (MI) reposent sur l'hypothèse qu'un modèle simple avec moyenne et variance communes est vérifié dans chaque cellule. L'étude de la robustesse des méthodes d'estimation de la variance est un aspect important de la simulation, parce qu'en pratique, les hypothèses qui sous-tendent les méthodes ne sont pour ainsi dire jamais entièrement satisfaites.

À la section suivante, nous décrivons brièvement trois méthodes d'estimation de la variance avec données imputées par la méthode hot deck. À la troisième section, nous décrivons la population étudiée, le plan d'échantillonnage utilisé dans les simulations et les méthodes appliquées pour générer les données manquantes et mettre en œuvre les imputations hot deck. À la quatrième section, nous donnons les résultats des simulations. Enfin, à la dernière section, nous présentons certaines conclusions concernant les méthodes et leur applicabilité.

2. Description des méthodes d'estimation de la variance

Nous représentons l'échantillon complet par A , le sous-ensemble qui répond à une question par A_R , et le sous-ensemble qui ne répond pas à la question par A_M . Pour les imputations, nous répartissons les unités en cellules hot deck portant l'indice $g = 1, \dots, G$, où le sous-ensemble de répondants dans la cellule g est A_{Rg} , et le sous-ensemble de non-répondants est A_{Mg} . Pour chaque unité pour laquelle une valeur manque, la méthode hot deck consiste à sélectionner aléatoirement dans la même cellule hot deck un répondant qui deviendra le donneur de la valeur imputée. Sous imputation hot deck, les donneurs sont souvent sélectionnés dans une cellule par échantillonnage aléatoire simple avec remise (casar), par échantillonnage aléatoire simple sans remise (cass) ou par échantillonnage avec probabilité proportionnelle au poids de sondage avec remise (eppsar). Puisque les résultats de la simulation obtenus en utilisant les méthodes casar et eppsar sont fort semblables, seuls les résultats pour la seconde-nommée hot deck pondérée-sont présentés ici. L'estimateur imputé d'un total de population est $\hat{\theta}_I = \sum_{i \in A_M} w_i^{-1} y_i + \sum_{i \in A_R} w_i^{-1} y_i^*$, où y_i^* est la valeur imputée pour l'unité i dans l'ensemble de non-répondants.

Estimation de la variance avec imputation hot deck : Une étude par simulation de trois méthodes

J. Michael Brick, Michael E. Jones, Graham Kalton et Richard Valliant¹

Résumé

Les méthodes d'estimation de la variance des estimations par sondage applicables à des données complètes sont biaisées lorsque certaines données sont imputées. Nous recourons à la simulation multiple pour estimer la variance d'un total par modèle, de la méthode du jackknife ajusté et de la méthode d'imputation multiple pour estimer la variance d'un total quand les réponses à certaines questions ont été imputées par la méthode hot deck. La simulation vise à étudier les propriétés des estimations de la variance des estimations imputées de totaux pour la population dans son ensemble et pour certains domaines provenant d'un plan d'échantillonnage stratifié non proportionnel à un degré quand les hypothèses sous-jacentes, comme l'absence de biais dans l'estimation ponctuelle et l'hypothèse des réponses manquantes au hasard dans les cellules hot deck, ne sont pas vérifiées. Les estimateurs de la variance des estimations pour l'ensemble de la population produisent des intervalles de confiance dont le taux de couverture s'approche du taux nominal, même en cas d'écarts modestes par rapport aux hypothèses, mais il n'en est pas ainsi des estimations par domaine. La couverture est surtout sensible au biais dans les estimations ponctuelles. Comme le démontre la simulation, même si une méthode d'imputation donne des estimations presque sans biais pour la population dans son ensemble, les estimations par domaine peuvent être fort biaisées.

Mots clés : Jackknife ajusté; estimation par domaine; estimation de la variance assistée par modèle; imputation multiple; non-réponse.

1. Introduction

L'imputation est un moyen fréquemment utilisé dans les recherches par sondage pour remplacer les réponses manquantes à certaines questions, de façon à produire des ensembles de données complets pour la diffusion au grand public ou l'analyse générale. Il est généralement reconnu que traiter des valeurs imputées comme s'il s'agissait de valeurs observées produit un biais par défaut dans les estimations de la variance des estimations par sondage. Par conséquent, le taux de couverture des intervalles de confiance est inférieur au taux nominal. Le biais dans les estimations de la variance a tendance à croître avec le taux de non-réponse partielle et peut être considérables si ce taux est élevé.

Nous étudions ici trois méthodes mises au point pour estimer la variance en présence de données imputées, à savoir une méthode assistée par modèle (Särndal 1992), une méthode du jackknife ajusté (Rao et Shao 1992) et une méthode d'imputation multiple (Rubin 1987). Chacune de ces méthodes a été évaluée théoriquement et par des méthodes de simulation, principalement sous des conditions concordant avec les hypothèses des méthodes. Dans la présente étude, nous utilisons la simulation pour comparer les trois méthodes dans des conditions expérimentales identiques sous lesquelles certaines hypothèses que requièrent les méthodes ne tiennent pas. L'objectif est d'examiner la

performance relative des méthodes dans des situations susceptibles de se produire en pratique. D'autres études en simulation des méthodes d'estimation de la variance avec données imputées ont généralement été plus limitées. Même l'étude par simulation de plus grande portée réalisée par Lee, Rancourt et Särndal (2001) était basée sur de petites populations et n'incluait pas l'imputation multiple.

Nous utilisons un échantillon stratifié non proportionnel à un degré, sélectionné à partir d'un ensemble de données de population réelles, pour évaluer ces méthodes d'estimation des valeurs imputées par la méthode hot deck, qui est l'une des méthodes d'imputation les plus populaires en recherche par sondage. Puisque l'imputation hot deck est une forme d'imputation par la régression (Kalton et Kasprzyk 1986), n'est pas une caractéristique critique en ce qui concerne l'étude des effets sur l'estimation de la variance. Nous étudions l'estimation pour les totaux de population, ainsi que pour les totaux de domaine. Dans le cas des estimations par domaine, nous supposons que l'indicateur de domaine est connu pour nous les membres de l'échantillon.

Dans les simulations, nous utilisons trois combinaisons distinctes de mécanismes de génération de données manquantes et de formation de cellules hot deck pour évaluer les propriétés des méthodes d'estimation de la variance sous des conditions qui violent à des degrés divers les hypothèses

- Rubin, D.B. (1987). *Multiple Imputation For Nonresponse In Surveys*. New York: John Wiley & Sons, Inc.
- Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 339-349.
- Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.
- Shao, J., Chen, Y. et Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.
- Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.
- Fuller, W.A. (1992). Variance estimation for sampling with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.
- Wang, N., et Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.
- Yung, W., et Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of American Statistical Association*, 95, 903-915.

7. Résumé

Dans l'imputation fractionnaire, plusieurs donneurs sont utilisés pour chaque valeur manquante et une fraction du poids du non-répondant est attribuée à chaque donneur. Si l'on utilise tous les donneurs, la procédure est entièrement efficace, sous le modèle, pour toutes les fonctions d'un vecteur y . Nous montrons que l'utilisation de l'imputation fractionnaire avec un petit nombre d'imputations par non-répondant peut donner un estimateur entièrement efficace de la moyenne. Les estimations d'autres paramètres, comme les estimations de la distribution cumulative sont presque entièrement efficaces.

L'imputation fractionnaire permet de construire des répliques d'usage général pour l'estimation de la variance. Il est possible d'utiliser un seul ensemble de répliques pour estimer la variance dans le cas de variables imputées, de variables observées sur l'ensemble des répondants et, sous les hypothèses du modèle, pour des fonctions de deux types de variables. Les répliques donnent des estimations des variances des moyennes de domaine dont le biais est nettement plus faible que celui des estimations par imputation multiple. Le biais tend vers zéro quand la valeur de M augmente et, dans la simulation, est modéré pour $M = 5$. L'estimateur de la variance par rééchantillonnage est facile à appliquer au moyen d'un logiciel de rééchantillonnage, tel que Wesvar.

L'imputation fractionnaire avec un nombre fixe de donneurs par receveur est un peu plus efficace pour la moyenne que l'imputation multiple avec le même nombre de donneurs. L'imputation fractionnaire donne des estimations de variance dont le biais est plus faible et dont la variance est nettement plus faible que les estimateurs par imputation multiple avec le même nombre d'imputations.

8. Remerciements

La présente étude a été financée partiellement aux termes d'un sous-contrat entre Westat et la Iowa State University en vertu du contrat n° ED-99-CO-0109 établi entre Westat et le Department of Education, ainsi que du contrat de coopération 13-3AEU-0-80064 conclu entre la Iowa State University, le U.S. National Agricultural Statistics Service et le U.S. Bureau of the Census. Nous remercions Jean Opsomer et Damiao Da Silva de leurs commentaires constructifs.

Bibliographie

Chen, J., Rao, J.N.K. et Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

Les hypothèses requises pour l'estimation de variance MI ne sont pas satisfaites pour l'ensemble de paramètres C. Par conséquent, la variance MI estimée est fortement biaisée pour tous les paramètres. Voir le tableau 6.5. Pour $M = 5$, le biais dans la variance MI estimée est d'environ 17 % pour la variance de la moyenne globale et de près de 50 % pour la moyenne de domaine. Le biais de la variance MI de la moyenne est plus faible pour une variable binomiale que pour une variable continue, parce que l'effet de stratification est plus faible dans le premier cas.

La variance de l'estimation de la variance MI est de 2,4 à 3,5 fois plus élevée que la variance de l'estimation de la variance FI pour $M = 5$ et de 3 à 7 fois plus élevée pour $M = 3$, ce qui démontre la supériorité nette de l'estimateur de variance FI pour cette configuration.

Tableau 6.4 Moyenne et variance des estimateurs ponctuels sous les conditions C (5 000 échantillons de taille 100)

Paramètre	Scénario d'imputation	Moyenne	Variance	Variance relative à FI (%)
(θ ₁)	Echantillon complet	2,10	0,00500	48
	FI(3)	2,10	0,01050	100
	ABB(3)	2,10	0,01220	116
(θ ₂)	Echantillon complet	2,01	0,02530	102
	FI(3)	2,01	0,02510	101
	ABB(3)	2,01	0,02850	115
Pr(Y < 2)	Echantillon complet	2,01	0,02710	109
	FI(5)	2,01	0,02480	100
	ABB(5)	2,01	0,02710	109
Pr(Y < 2)	Echantillon complet	0,45	0,00127	45
	FI(3)	0,45	0,00281	100
	ABB(3)	0,45	0,00322	115
(θ ₃)	FI(5)	0,45	0,00280	100
	ABB(5)	0,45	0,00314	112
Pr(Y < 1)	Echantillon complet	0,15	0,00107	54
	FI(3)	0,15	0,00199	100
	ABB(3)	0,15	0,00226	114
(θ ₄)	FI(5)	0,15	0,00199	100
	ABB(5)	0,15	0,00214	108

Moyenne relative, statistique t et variance relative pour les estimateurs de variance sous les conditions C (5 000 échantillons de taille 100)

Paramètre	Méthode	Moyenne relative (%)	Statistique t*	Variance relative (%)
(θ ₁)	FI(3)	100,9	0,41	6,42
	ABB(3)	116,7	7,31	40,14
	FI(5)	100,8	0,39	6,42
Moyenne de domaine	ABB(5)	117,1	7,99	22,29
	FI(3)	122,7	10,78	16,23
	ABB(3)	144,4	19,79	46,05
Pr(Y < 2)	FI(5)	106,1	2,95	11,95
	ABB(5)	148,7	22,51	32,49
	FI(3)	104,4	2,18	6,63
(θ ₃)	ABB(3)	114,7	6,54	42,32
	FI(5)	101,8	0,89	6,42
	ABB(5)	112,1	5,74	20,67
Pr(Y < 1)	FI(3)	102,3	1,13	11,08
	ABB(3)	101,3	0,58	39,14
	FI(5)	99,9	-0,04	10,05
ABB(5)		102,2	1,04	23,60

* Statistique pour l'hypothèse selon laquelle la variance estimée est sans biais.

Moyenne et variance des estimateurs ponctuels sous les conditions A (5 000 échantillons de taille 100)

Tableau 6.2

Paramètre	Schéma d'imputation	Moyenne	Variance	Variance relative à FI (%)
(θ_1)	FI(3)	1,00	0,00849	100
	ABB(3)	1,00	0,00926	109
	FI(5)	1,00	0,00849	100
	ABB(5)	1,00	0,00903	106
Moyenne de domaine	Echantillon complet	1,14	0,02020	99
(θ_2)	FI(3)	1,14	0,02050	100
	ABB(3)	1,14	0,02230	109
	FI(5)	1,14	0,02040	100
	ABB(5)	1,14	0,02170	106
Pr($Y < 2$)	Echantillon complet	0,87	0,00104	51
(θ_3)	FI(3)	0,87	0,00202	100
	ABB(3)	0,87	0,00228	113
	FI(5)	0,87	0,00202	100
	ABB(5)	0,87	0,00223	110
Pr($Y < 1$)	Echantillon complet	0,50	0,00208	66
(θ_4)	FI(3)	0,50	0,00313	100
	ABB(3)	0,50	0,00342	109
	FI(5)	0,50	0,00313	100
	ABB(5)	0,50	0,00329	105

Moyenne relative, statistique t et variance relative pour les estimateurs de variance sous les conditions A (5 000 échantillons de taille 100)

Tableau 6.3

Paramètre	Méthode	Moyenne relative (%)**	Statistique t^*	Variance relative (%)
Moyenne	FI(3)	100,1	0,05	5,66
(θ_1)	ABB(3)	99,6	-0,19	19,25
	FI(5)	100,1	0,03	5,65
	ABB(5)	98,2	-0,89	9,95
Moyenne de domaine	FI(3)	115,9	7,54	13,88
(θ_2)	ABB(3)	127,9	12,72	28,88
	FI(5)	106,6	3,14	11,62
	ABB(5)	128,4	13,43	20,03
Pr($Y < 2$)	FI(3)	103,9	1,86	13,90
(θ_3)	ABB(3)	100,8	0,36	48,42
	FI(5)	101,7	0,82	12,07
	ABB(5)	98,5	-0,67	25,10
Pr($Y < 1$)	FI(3)	98,5	-0,75	4,67
(θ_4)	ABB(3)	96,3	-1,80	18,51
	FI(5)	97,6	-1,20	4,45
	ABB(5)	96,7	-1,65	10,17

* Statistique pour l'hypothèse selon laquelle la variance estimée est sans biais.

** Moyenne de Monte-Carlo des estimations de variance divisée par la variance de Monte-Carlo des estimations, en pourcentage.

Dans un deuxième ensemble de paramètres, noté C, les moyennes étaient les suivantes :

Cellule 1 des strates 1 à 25; $\mu = 0,4$

Cellule 1 des strates 26 à 50; $\mu = 3,0$

Cellule 2 des strates 1 à 25; $\mu = 1,6$

Cellule 2 des strates 26 à 50; $\mu = 2,2$.

Tous les autres paramètres sont les mêmes que dans l'ensemble de paramètres A. Les propriétés des estimateurs sont données au tableau 6.4. L'imputation fractionnaire (FI) et l'imputation multiple (MI) produisent toutes deux des estimations sans biais des moyennes et de la moyenne de domaine. Comme pour l'ensemble de paramètres A, la procédure FI est de 8 % à 12 % plus efficace que la procédure MI pour $M = 5$ et de 14 % à 16 % plus efficace pour $M = 3$.

bootstrap bayésien approximatif (ABB) de Rubin et Schenker (1986) avec $M = 5$ et avec $M = 3$. Les procédures FI et MI sont toutes deux sans biais pour les quatre paramètres du tableau 6.2. La dernière colonne de ce tableau donne la variance de Monte-Carlo de l'estimateur divisée par la variance de Monte-Carlo de la procédure FI avec $M = 5$, exprimée en pourcentage. La procédure FI est de 5 % à 10 % plus efficace que la procédure MI avec $M = 5$ et de 9 % à 13 % plus efficace que la procédure MI avec $M = 3$.

Sous le modèle, la moyenne des valeurs observées n'est pas le meilleur estimateur de la moyenne de domaine. Dans cet exemple, l'estimateur FI est presque aussi efficace que l'estimateur pour l'échantillon complet. L'effet d'un nombre plus petit d'observations est compensé par l'utilisation d'un meilleur estimateur de la moyenne pour le domaine. Sous le modèle, l'indicateur de domaine est indépendant des valeurs de y , sachant la cellule. Par conséquent, il est efficace d'utiliser toutes les valeurs contenues dans la cellule comme donneurs, plutôt que simplement les répondants dans le domaine.

Les propriétés des estimateurs de la variance sont données au tableau 6.3. La colonne intitulée « moyenne relative » donne la moyenne estimée de Monte-Carlo des variances estimées divisées par la variance estimée de Monte-Carlo, où cette dernière est donnée au tableau 6.2. Les deux méthodes d'estimation de la variance semblent être quasiment sans biais pour la variance de la moyenne. La variance relative de l'estimateur de variance MI pour $M = 5$ est égale à près de deux fois celle de l'estimateur de variance FI pour $M = 5$. Pour $M = 3$, l'estimateur de variance MI vaut plus de trois fois l'estimateur de variance FI. La variance de l'estimateur de variance MI est grande, parce que la variance due aux observations manquantes est estimée avec quatre degrés de liberté pour $M = 5$ et avec deux degrés de liberté pour $M = 3$.

L'estimateur de variance MI de la moyenne de domaine est gravement biaisé. Cette propriété a été reconnue pour la première fois par Fay (1991, 1992) et étudiée par Meng (1994), ainsi que par Wang et Robins (1998). L'estimateur de variance FI pour la moyenne de domaine présente aussi un biais positif, quoique nettement plus faible que celui de MI. Nous pouvons réduire le biais dans l'estimateur de variance FI en augmentant M , mais le biais de MI dépend peu de M .

Tous les estimateurs de variance de θ_4 présentent un léger biais négatif. Nous pensons que l'estimateur FI est légèrement biaisé pour θ_4 parce que, bien que nous utilisions le vecteur z , les poids sont légèrement biaisés par la procédure de régression. Il est connu que l'imputation multiple (MI) donne lieu à un biais de petit échantillon. Voir Kim (2002).

6.3 Résultats de l'étude de Monte-Carlo

Les résultats de Monte-Carlo pour les 5 000 échantillons générés par les paramètres du tableau 6.1 sont donnés aux tableaux 6.2 et 6.3. Nous présentons les résultats pour l'échantillon complet, pour l'imputation fractionnaire avec cinq donneurs, pour l'imputation fractionnaire avec trois donneurs et pour l'imputation multiple (MI) en utilisant le

L'estimateur entièrement efficace (FE) avec unité supprimée pour la réplique k de la moyenne de cellule de z est défini en (19). Le poids fractionnaire initial du donneur k à l'élément j est fixé à $w_{jk}^{(0)} = 0,01w_{jk}^*$. Ce poids initial assure que le poids final soit faible, mais permet l'ajustement par la régression. Les poids finaux $w_{jk}^{(t)}$ sont calculés par la procédure de régression (18) en utilisant le poids initial $w_{jk}^{(0)}$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Le vecteur z est l'estimateur entièrement efficace de la moyenne pour chacune des cinq variables comprises dans le l'estimateur par imputation, modifié par la régression, de la pour $s = 1, 2, 3, 4$ et soit $z_{gi} = (z_{g1i}, z_{g2i}, \dots, z_{g5i})$.

Tableau 5.4
Répliques jackknife de la moyenne de cellule des variables nominales de la variable x

Cellule	Niveau de x				Réplique			
1	1	0,33	0,67	0,50	0,33	0,50	0,33	0,50
	2	0,00	0,33	0,33	0,25	0,33	0,25	0,25
	3	0,33	0,00	0,00	0,25	0,00	0,25	0,00
	1	0,00	0,00	0,50	0,00	0,00	0,00	0,50
	2	0,50	0,50	0,50	0,33	0,33	0,67	0,50
2	3	0,50	0,50	0,50	0,67	0,33	0,33	0,50
	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	2	0,50	0,50	0,50	0,67	0,67	0,67	0,33
	1	0,00	0,33	0,33	0,00	0,25	0,25	0,25
	2	0,50	0,50	0,50	0,33	0,33	0,67	0,50

Tableau 5.5
Poids jackknife pour l'imputation fractionnaire

Obs.		Réplique									
1	0	1,1111	1,1111	1,1111	0,4205	0,3206	0,4563	1,1111	1,1111	1,1111	1,1111
2	0,1664	0	0,4400	0,3002	0,4400	0,2500	0,4563	0,3206	0,4400	0,3206	0,2724
2	0,6559	0	0,3706	0,3904	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
2	0,8888	0,3706	0	0,3706	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
3	0,3706	0,3706	0	0,3706	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
3	0,3697	0,3697	0	0,3697	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
3	0,3708	0,3708	0	0,3708	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
4	0,3704	0,3704	0	0,3704	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
4	0,3704	0,3704	0	0,3704	0,3505	0,4048	0,3505	0,3179	0,3505	0,4400	0,5075
5	1,1111	1,1111	1,1111	1,1111	0	0,2778	0,2778	0,2778	0,2778	0,2778	0,2778
6	1,1111	1,1111	1,1111	1,1111	0	0,2778	0,2778	0,2778	0,2778	0,2778	0,2778
7	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111
8	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111
9	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111
10	0	0,4623	0,2778	0,2324	0,2778	0,3008	0,2220	0,2957	0,4235	0	0

6. Études par simulation

6.1 Paramètres d'intérêt

Pour étudier les propriétés de la méthode d'imputation, nous avons réalisé une étude de Monte-Carlo. L'échantillon est stratifié, avec deux éléments par strate et deux cellules d'imputation, où les cellules recourent les strates. La cellule 1 comprend 20 % de la population des strates 1 à 25 et 80 % de la population des strates 26 à 50. La probabilité de réponse est 0,7 pour la cellule 1 et 0,5 pour la cellule 2. Nous examinons deux variables. La variable D est toujours observée et définit une sous-population. La probabilité que $D = 1$ est de 0,25 pour la cellule 1 et de 0,40 pour la cellule 2. La variable y est sujette à la non-réponse avec probabilités de réponse dans les cellules constantes. La variable D est indépendante de y et de la probabilité de réponse. La variable y suit une loi normale, où les paramètres pour une population de 50 strates sont donnés au tableau 5.1. Dans le modèle de génération des données du tableau 6.1, il n'existe aucun effet de strate. Les paramètres d'intérêt sont : θ_1 = moyenne de y , θ_2 = moyenne de y

6.2 Méthodes d'estimation

Dans la simulation, nous avons utilisé $M = 5$ et $M = 3$ donneurs par receveur. Nous avons sélectionné des échantillons systématiques à titre de donneurs pour chaque receveur. Si le nombre de répondants dans la cellule est inférieur à M , chaque répondant est utilisé comme donneur pour chaque receveur et les w_{ij}^h sont proportionnels au poids original des répondants. Si le nombre de répondants dans la cellule est supérieur à M , nous classons les donneurs par taille et les numérotions de 1 à r^g . Puis, nous plaçons les donneurs dans l'ordre 3, 5, ..., r^g , r^{g-1} , r^{g-3} , ..., 2 pour les

Ensemble de paramètres A		Poids de		Strates l'élément	
Cellule 1	Cellule 2	Moyenne	Variance	Moyenne	Variance
1,6	0,36	0,4	0,36	0,01	0,01
1,6	0,36	0,4	0,36	0,01	0,01
26 à 50	1 à 25	0,01	0,01	0,01	0,01

Tableau 5.2
Poids fractionnaires pour les moyennes

Observation	Poids	Donneur	Cellule	Cellule	x	y
1	1,0000	1	1	1	1	7
2	0,2886	1	1	1	1	7
2	0,3660	6	1	1	1	15
2	0,3154	8	1	1	1	9
3	0,3333	5	1	1	1	3
3	0,3334	7	1	1	1	8
3	0,3334	9	1	1	1	2
4	0,5000	1	1	1	1	14
4	0,2500	1	1	1	1	14
4	0,2500	1	1	1	1	14
5	1,0000	1	1	1	1	3
5	1,0000	2	1	1	1	15
6	1,0000	2	2	2	2	15
7	1,0000	2	2	2	2	8
8	1,0000	1	3	3	3	9
8	1,0000	2	2	2	2	8
9	1,0000	2	2	2	2	2
10	0,2247	8	2	2	2	9
10	0,2753	4	2	2	1	14
10	0,2905	1	1	1	1	7
10	0,2905	6	2	2	1	15

Nous attribuons une fraction initiale égale à un tiers aux trois valeurs imputées pour les observations 3 et 4, et une fraction initiale égale à un quart aux quatre valeurs imputées pour l'observation 10. Puis, nous ajustons les poids fractionnaires en utilisant la méthode de régression de l'équation (18) pour donner la moyenne par imputation fractionnaire entièrement efficace (FEFI) de y comme estimateur, où l'estimateur entièrement efficace de la moyenne de y est

$$\hat{y}_{FE} = \sum_{k=1}^8 \frac{n_k}{n} \hat{y}_{k8} = 8,4833.$$

Nous contrainsons les poids pour l'observation 10 de sorte que les fractions estimées pour les deux catégories de x soient les fractions de cellule. Alors, comme la moyenne pondérée de la variable nominale est contrôlée pour chaque individu, le vecteur z contient uniquement la variable y . Le tableau 5.2 donne les poids fractionnaires finaux calculés sous pondération par la régression. Un analyste peut utiliser l'ensemble de données du tableau 5.2 et tout programme informatique pour échantillon complet pour calculer des estimations des fonctions de y et x , telles que la moyenne de y pour les catégories de x . L'ensemble de données fractionnaires est entièrement efficace pour toute fonction de la variable x et est également entièrement efficace pour la moyenne de la variable y .

Tableau 5.3
Répliques jackknife de la moyenne de cellule de la variable y

Cellule	1	2	3	4	5	6	7	8	9	10
Réplique	1	2	3	4	5	6	7	8	9	10
	12,67	4,33	11,25	10,33	11,25	10,00	11,25	12,00	11,25	11,25
	4,33	4,33	4,33	4,33	5,00	4,33	2,50	4,33	5,50	4,33

Pour l'estimation de la variance par le jackknife, nous répétons le calcul des poids pour chaque réplique. Les estimations répétées des moyennes de cellule de y sont données au tableau 5.3 et les estimations répétées des fractions pour les catégories de x sont données au tableau 5.4. Nous utilisons les valeurs des tableaux 5.3 et 5.4 comme totaux de contrôle $\bar{z}_{FE,k}$ dans la pondération par la régression. Nous prenons $w_{jk}^{(k)} = 3^{-1}$ comme valeur initiale des fractions de rééchantillonnage pour l'observation 2 et $w_{jk}^{(k)} = 4^{-1}$ pour l'observation 10.

Le tableau 5.5 contient les poids jackknife pour l'ensemble de données obtenu par imputation fractionnaire du tableau 5.2. Les poids de rééchantillonnage sont utilisés de la même façon que les répliques pour un échantillon complet. Ils conviennent, avec les mises en garde de la section suivante, pour toute statistique pour laquelle le jackknife avec échantillon complet est approprié. Donc, la procédure est particulièrement séduisante pour un ensemble de données d'usage général, car l'analyste ne doit effectuer aucun calcul supplémentaire.

Nous obtenons l'estimateur entièrement efficace de la moyenne de y en considérant que les répondants représentent la deuxième phase d'un échantillon à deux phases. Un estimateur de variance pour échantillon à deux phases peut s'écrire

$$V = \frac{1}{2} \sum_{k=1}^8 \frac{n_k}{n} (\hat{y}_{k8} - \hat{y}_{FE})^2 + \sum_{k=2}^8 \left(\frac{n_k}{n} \right)^2 \frac{1}{s_{k8}^2} = 3,043,$$

où s_{k8}^2 est la variance d'échantillon intracellulaire pour la cellule k . Si nous utilisons les poids de rééchantillonnage du tableau 5.5, l'estimation de la variance par rééchantillon-nage pour la moyenne de y est

$$\hat{V}_{jk}(\hat{y}_{PI}) = \sum_{k=10}^8 0,9 (\hat{y}_{PI}^{(k)} - \bar{y}_{PI})^2 = 3,078.$$

La différence entre l'estimateur de la variance linéarisé et l'estimateur de la variance par le jackknife est

$$\sum_{k=1}^8 \left(\frac{r_k}{r_k - 1} \frac{n - 1}{n} - 1 \right) \frac{1}{s_{k8}^2}.$$

Donc, l'estimateur de la variance par le jackknife surestime légèrement la variance réelle dans notre exemple.

Soit $\mathbf{z}_{FE, g}$ l'estimateur entièrement efficace pour la cellule g . Si nous utilisons des procédures de régression, les w_i modifiés pour donner la moyenne de cellule entièrement efficace de \mathbf{z} , sont

$$(81) \quad \left(\left(\mathbf{z}^{(f, g)} - {}^{f[t]} \mathbf{z}^{(g)} \right) {}^0 \hat{w}_*^g \mathbf{M}^g \mathbf{S} \left(\mathbf{z}^{(g)} - \mathbf{z}^{(f, \text{FE}, g)} \right) + {}^0 \hat{w}_*^0 \mathbf{M} \right) = \hat{w}_*^0 \mathbf{M}$$

$$\begin{aligned} & \text{, } \overset{\text{f}}{\text{f}} \text{p}^{\text{f}[\text{f}]^{\text{g}}} \text{z}^{\text{0f}} \overset{\text{f}}{\text{f}} \text{m} \overset{\text{f}}{\text{f}} \text{y} \text{z}^{\text{f}} \overset{\text{f}}{\text{f}} \text{q} \overset{\text{f}}{\text{f}} \text{z} = \overset{\text{f}}{\text{f}} \text{z} \\ & \text{, } \overset{\text{f}}{\text{f}} \text{p}^{\text{f}[\text{f}]^{\text{g}}} \text{z}^{\text{0f}} \overset{\text{f}}{\text{f}} \text{m} \overset{\text{f}}{\text{f}} \text{z} = \text{f} \cdot \overset{\text{f}}{\text{f}} \text{z} \\ & \text{, } \overset{\text{f}}{\text{f}} \text{p}(\text{f} \cdot \overset{\text{f}}{\text{f}} \text{z} - \text{f}[\text{f}]^{\text{g}} \text{z}) \text{, } (\text{f} \cdot \overset{\text{f}}{\text{f}} \text{z} - \text{f}[\text{f}]^{\text{g}} \text{z})^{\text{0f}} \overset{\text{f}}{\text{f}} \text{m} \overset{\text{f}}{\text{f}} \text{y} \text{z}^{\text{f}} \overset{\text{f}}{\text{f}} \text{q} \overset{\text{f}}{\text{f}} \text{z} = \overset{\text{f}}{\text{f}} \text{z} \text{z} \end{aligned}$$

$$\frac{1}{L} \left(\sum_{s \in \Lambda_{Lg}} w_s^f \right) = q^f$$

Pour estimer la variance, nous créons des réplicates de sorte que les poids appliqués aux données reflètent l'effet de la suppression d'un élément sur l'estimateur entièrement efficace. Nous utilisons les mots « suppression de ce » et « supprimer » pour identifier l'élément choisi pour la modification principale du poids pour l'estimation de la variance par

Soit $w_i^{(k)}$ le poids attribué à l'élément i pour la k^{e} répétition de l'estimation de la variance de l'estimateur pour l'échantillon complet. Alors, la réplique pour la moyenne entièrement efficace de y pour la cellule g est

$$(61) \quad \sum_{i \in A_{Rg}} M_{(k)}^i \left[\sum_{i \in A_{Rg}} M_{(k)}^i \right]^{-1} = \underline{z}_{(k)}^g$$

Les fractions de réchantillonnage sont attribuées aux donneurs dans la cellule g de sorte que l'estimation de la moyenne de cellule par réchantillonnage soit $\bar{\mathbf{z}}^{(k)}_g$. Nous assignons les poids fractionnaires initiaux $w^{(k)}_{*0}$ ou $w^{(k)}_{*0}$ assez faible, mais positif, si k est une unité supprimée pour la réplique k . Nous calculons les poids fractionnaires finaux $w^{(k)}_{*1}$ selon la procédure (18) en remplaçant $\mathbf{z}^{\text{FE},g}$ par $\mathbf{z}^{(k)}_g$ et $w^{(k)}_{*0}$ par $w^{(k)}_{*0}$. La procédure simule l'effet de la suppression d'un seul élément sur l'estimateur entièrement efficace.

5. Un exemple artificiel

Nous présentons ici un exemple fondé sur des données artificielles afin d'illustrer l'application de la méthode

proposée. Supposons que nous observons deux variables d'intérêt, x et y , dans un échantillon de taille $n = 10$ obtenu par échantillonnage aléatoire simple. La variable x est un variable nominale comptant trois catégories, disons 1, 2 et 3, et la variable y est une variable continue. Il y a non-réponse partielle pour les deux variables et il existe un ensemble de cellules d'imputation pour chaque variable. Le tableau 5.1 donne les observations sur l'échantillon, où la non-réponse est représentée par M . Nous utilisons un poids unitaire pour simplifier la présentation. Nous divisons par dix pour obtenir les poids pour la moyenne.

Tableau 5.1

Observation	Poids	Cellule	pour x	Cellule	x	y
1	1	1	1	1	1	7
2	1	1	1	1	2	M
3	1	1	1	2	3	M
4	1	1	1	1	14	M
5	1	1	1	2	1	3
6	1	1	2	1	2	15
7	1	1	2	2	3	8
8	1	1	2	1	3	9
9	1	1	2	2	2	2
10	1	1	2	1	2	M

Comme la variable x est une variable nominale à trois catégories, l'utilisation de trois fractions pour l'imputation nominale donne des estimations entièrement efficaces pour la distribution de la variable x . Donc, dans le tableau 5.2, les poids pour les trois valeurs imputées de x pour la quatrième observation sont les fractions pour les trois catégories dans la cellule 1 pour x .

Si l'on utilise un sous-ensemble de donneurs pour catégoriser dans la cellule l pour x .

Dans la situation où les réponses à deux questions manquent, plusieurs approches sont possibles, y compris la formation d'un troisième ensemble de cellules d'imputation pour ce genre de cas. Étant donné la petite taille de l'échantillon dans notre illustration, nous imputons sous l'hypothèse que x et y sont indépendantes dans les cellules. Donc, chacune des deux valeurs possibles de x , nous imputons deux valeurs de y . Nous choisissons l'une des imputures à la moyenne des réponses et l'autre, de façon à ce qu'elle soit plus grande que la moyenne. Voir les valeurs imputées pour l'observation 10 au tableau 5.2.

procédure comportant un nombre fixe de donneurs par receveur qui est entièrement efficace pour le total général, mais qui n'est pas forcément entièrement efficace pour les sous-populations. La méthode consiste à affecter des donneurs pour produire une variance faible de la pondération des imputées entre receveurs et à modifier la pondération des donneurs pour arriver à l'efficacité complète pour le total.

Supposons que M donneurs soient affectés à chaque receveur. Nous proposons d'affecter les donneurs aux receveurs de façon à approximer la distribution de tous les répondants dans la cellule. L'une des méthodes de sélection possibles consiste à tirer un échantillon stratifié pour chaque receveur. Une autre consiste à recourir à l'échantillonnage systématique avec probabilités proportionnelles aux poids pour sélectionner les donneurs pour chaque receveur. Les fractions initiales w_{j0}^* sont affectées aux valeurs données. Dans le cas de l'échantillonnage systématique avec poids égaux, la fraction initiale w_{j0}^* est M^{-1} .

Après avoir affecté les donneurs, nous corrigeons les fractions initiales, w_{j0}^* , de sorte que la somme des poids donne l'estimateur entièrement efficace de la moyenne de y et que la fonction de distribution cumulative estimée d'après les poids soit une approximation de l'estimateur entièrement efficace de la fonction de distribution cumulative. La modification de la pondération par la régression a été proposée par Fuller (1984, 2003). Chen, Rao et Sitter (2000) discutent d'une méthode d'imputation efficace où l'on modifie les valeurs imputées plutôt que les poids. Soit $\mathbf{z}^{(g)} =$

$$\begin{aligned} z_{g1}^* &= y_j \\ z_{g2}^* &= 1 \text{ si } y_j \leq L_2 \\ &= 0 \text{ autrement} \\ &\vdots \\ z_{gj}^* &= 1 \text{ si } L_{\alpha-1} < y_j \leq L_\alpha \\ &= 0 \text{ autrement} \end{aligned}$$

où $L_2, L_3, \dots, L_\alpha$ divisent la fourchette de valeurs observées de y dans la cellule g en $\alpha - 1$ sections. Le nombre de sections que l'on peut utiliser dépend du nombre et du type d'observations dans la cellule, du nombre de receveurs et du nombre de donneurs par receveur. Si le nombre de donneurs par receveur est grand, il est possible d'ajuster l'ensemble de poids pour chaque receveur de façon à ce que la somme des w_{j0}^* sur j soit égale à l'unité pour chaque j et que la somme des $w_{j0}^* y_j$ sur j soit l'estimateur entièrement efficace pour chaque j . Dans la plupart des cas, les poids sont ajustés de sorte que la somme des w_{j0}^* sur j soit égale à l'unité pour chaque j et que les moyennes de cellule des valeurs imputées soient égales à l'estimateur entièrement efficace.

La méthode proposée est étroitement associée à l'estimateur de Rao et Shao (1992). Voir aussi Yung et Rao (2000). Toutefois, l'utilisation de l'imputation fractionnaire simplifie beaucoup l'estimation de la variance. Dans la création des répliques, seuls les poids appliqués aux valeurs imputées changent. Il n'est pas nécessaire de recalculer les valeurs imputées et, une fois qu'ils sont calculés, les poids des répliques peuvent être utilisés pour n'importe quelle fonction lisse du vecteur y . En outre, les répliques fractionnaires rendent l'estimateur (16) approprié pour un vecteur de variables y .

Nous pouvons utiliser le théorème 3.1 de Kim, Navarro et Fuller (2005) pour montrer que, étant donné une méthode de production de répliques de l'échantillon complet convergente,

$$\hat{V}^{\text{HFI}}(\hat{\theta}^{\text{HFI}} | F_y) = -N_v^{-2} \sum_{g=1}^{G_v} \sum_{i \in U_{g_v}} \pi_{i-1}^{-1} (1 - \pi_{g_v}) e_2^i + o_p(n_v^{-1}), \quad (17)$$

où $\hat{\theta}^{\text{HFI}}$ est défini dans (10), et où la loi a trait aux mécanismes d'échantillonnage et de réponse.

Si l'on peut ignorer la correction pour population finie, l'estimateur (16) est convergent pour $V\{\theta^{\text{FE}}\}$. Si la taille d'échantillon est grande comparativement à N_v , alors un estimateur de

devrait être ajouté à (16). La méthode d'imputation et d'estimation de la variance décrite pour le modèle de réponses produit aussi des estimateurs convergents pour le modèle de moyenne de cellule. Sous ce modèle, les éléments contenus dans une cellule de la population finie sont une réalisation de variables aléatoires indépendantes et de même loi. La méthode d'imputation fondée sur le modèle de réponse n'est pas nécessairement entièrement efficace pour la moyenne de population sous le modèle de moyenne de cellule, mais on peut montrer que l'estimateur de la moyenne et l'estimateur de la variance de la moyenne estimée sont convergents.

4. Approximations de la méthode entièrement efficace

Aux sections précédentes, nous avons construit l'estimateur $\hat{\theta}^{\text{HFI}}$ de façon à ce que la variance due à l'imputation soit nulle. L'application de la méthode d'imputation fractionnaire, telle qu'elle est décrite en (11), pourrait nécessiter l'utilisation d'un grand nombre de donneurs pour chaque receveur. Par conséquent, nous décrivons une

que le nombre de répondants, r_g , soit nul. Quand cela se

produit en pratique, les cellules sont regroupées.

Nous pouvons obtenir les propriétés de grand échantillon

de l'estimateur pour une série de populations et d'échan-

tilons. Supposons que la population soit composée de G_v

cellules disjointes et exhaustives, où v est l'indice de la

série. Supposons que la variance d'un estimateur de la

moyenne pour l'échantillon sélectionné soit $O(n_v^{-1})$, où n_v est

la taille de l'échantillon sélectionné à partir de la v^e

population. Supposons que les réponses soient indépendantes.

Alors, sous des conditions de régularité, nous pouvons nous

servir des procédures utilisées par Kim, Navarro et Fuller

(2005) dans la preuve de leur théorème 2.1 pour montrer

que l'estimateur (7) satisfait

$$\hat{\theta}_{FE_v} = \hat{\theta}_v + \sum_{g_v=1}^{g_v} w_{g_v} (\pi_{g_v}^{-1} R_{g_v} - I) e_{g_v} + o_p(n_v^{-1/2} N_v), \quad (9)$$

où $e_{g_v} = y_{g_v} - \bar{y}_{g_v}$, A_{g_v} est l'ensemble d'indices d'échan-

tilon dans la g_v cellule pour le v^e échantillon, \bar{y}_{g_v} est la

moyenne de population de la variable y dans la cellule g_v

de population, F_v , π_{g_v} est la probabilité qu'un élément dans

la cellule g_v réponde, et F_v représente la v^e population.

En outre

$$V(\hat{\theta}_v | F_v) = V(\hat{\theta}_v | F_v)$$

$$+ E \left\{ \sum_{g_v=1}^{g_v} \pi_{g_v}^{-1} (1 - \pi_{g_v}) \sum_{i \in A_{g_v}} w_i^2 e_{iv}^2 | F_v \right\}, \quad (10)$$

où

$$\tilde{\theta}_{FE_v} = \hat{\theta}_v + \sum_{G_v} w_{g_v} (\pi_{g_v}^{-1} R_{g_v} - I) e_{g_v}.$$

Nous pouvons appliquer l'estimateur (7) en utilisant une

imputation fractionnaire dans laquelle chaque unité ré-

pondante figurant dans une cellule d'imputation est utilisée

comme donneur pour chaque non-répondant compris dans

la cellule. Alors, l'estimateur (7) peut s'écrire sous la forme

$$\hat{\theta}^{FEH} = \sum_{G_v} \sum_{j \in A_v \cap U_g} w_j w_{j^*} y_{j^*}^i, \quad (11)$$

où $w_j w_{j^*}$ est le poids du donneur i pour le receveur j , w_{j^*}

est la fraction d'imputation du donneur i pour le receveur j

définie dans (3), et

$$w_{j^*}^i = \begin{cases} 1 & \text{si } j \in A_v \cap U_g \\ w_j w_{j^*} & \text{si } j \in A_v \cap U_g \\ 0 & \text{si } j \in A_v \cap U_g \end{cases} \quad (12)$$

L'estimateur (11) avec w_{j^*} donné par (12), qui est

aléatoirement équivalent à (7), est appelé *estimateur par*

imputation entièrement efficace (FEHI pour *fully efficient*

fractionally imputed). L'estimateur par imputation fraction-

naire a l'avantage de permettre d'estimer directement des

fonctions de y , telles que la fraction inférieure à un nombre

donné, d'après l'ensemble de données imputées fraction-

naires.

Afin d'examiner l'estimation de la variance par rééchan-

tillonnage, posons qu'un estimateur de la variance par ré-

échantillonnage pour l'échantillon complet est donné par

$$V(\hat{\theta}) = \sum_{k=1}^L c_k^2 (\hat{\theta}^{(k)} - \hat{\theta})^2, \quad (13)$$

où $\hat{\theta}^{(k)}$ est la k^e estimation de θ , d'après les observations

incluses dans la k^e réplique, L est le nombre de répliques, et

c_k est un facteur associé à la réplique k déterminé par la

méthode de rééchantillonnage. Pour une discussion de la

répétition des échantillons d'enquête, voir Krewski et Rao

(1981), ainsi que Rao, Wu et Yue (1992). Si l'estimateur

original $\hat{\theta}$ est un estimateur linéaire de la forme (1), la k^e

estimation répétée de $\hat{\theta}$ peut s'écrire

$$\hat{\theta}^{(k)} = \sum_{G_v} w_{g_v}^i \left(\sum_{j \in A_v \cap U_g} w_{j^*}^i y_{j^*}^i \right) \quad (14)$$

où $w_{j^*}^i$ est le poids de rééchantillonnage de la i^e unité de

la k^e réplique.

Nous proposons pour l'estimateur $\hat{\theta}^{FEH}$ la réplique

$$V^{FEH} = \sum_{k=1}^L c_k^2 (\hat{\theta}^{(k)} - \hat{\theta}^{FEH})^2. \quad (16)$$

Les répliques données par (15) peuvent être calculées en

deux étapes. Premièrement, nous créons la réplique habi-

tuelle en définissant les poids $w_{j^*}^i$ pour chaque élément.

Deuxièmement, pour un non-répondant, nous utilisons

comme fraction d'imputation par rééchantillonnage du

donneur i au receveur j

$$w_{j^*}^i = \frac{\sum_{j \in A_v \cap U_g} w_{j^*}^i}{w_{j^*}^i}.$$

Notons que la somme des poids de rééchantillonnage

fractionnaire des enregistrements donneurs pour chaque

receveur est égale au poids de rééchantillonnage de chaque

unité dans un échantillon complet.

2. Conditions de base

Considérons une population de N éléments identifiés par un ensemble d'indices $U = \{1, 2, \dots, N\}$. À chaque unité i de la population est associée une variable étudiée y_i et un vecteur \mathbf{x}_i de données auxiliaires. L'ensemble de vecteurs, (y_i, \mathbf{x}_i) , $i = 1, 2, \dots, N$, est noté F .

Soit A les indices des éléments d'un échantillon sélectionné d'après un ensemble de règles probabilistes appelées *mécanisme d'échantillonnage*. Soit θ_N la quantité d'intérêt dans la population et θ un estimateur de θ_N , pour l'échantillon complet, linéaire en y et écrits que

$$(1) \quad \theta = \sum_{i \in A} w_i y_i.$$

Si w_i est l'inverse de la probabilité de sélection, alors θ est sans biais pour le total de population.

Soit A_R et A_M les ensembles d'indices pour les répondants et les non-répondants dans l'échantillon, respectivement. Définissons la fonction indicatrice de réponse

$$(2) \quad R_i = \begin{cases} 1 & \text{si } i \in A_R \\ 0 & \text{si } i \in A_M \end{cases}$$

et posons que $\mathbf{R} = \{(i, R_i); i \in A\}$. La loi de \mathbf{R} est appelée *mécanisme de réponse*.

Supposons que la population finie U soit constituée de G cellules d'imputation, où l'ensemble d'éléments dans la cellule g est U_g . Soit n_g le nombre d'éléments de l'échantillon compris dans la cellule d'imputation g et soit $r_g, r_g^* > 0$, le nombre de répondants dans la cellule d'imputation g . Supposons que nous ayons le modèle de réponse uniforme dans les cellules, où les r_g réponses dans une cellule sont équivalentes à un échantillon de Poisson tiré avec probabilités égales à partir des n_g éléments.

L'imputation fractionnaire est une méthode consistant à utiliser plus d'un donneur par receveur. Kalton et Kish (1984) ont proposé l'imputation fractionnaire comme méthode d'imputation efficace. Elle a été discutée par Fay (1996). Soit d_{ij} le nombre de fois que y_i est utilisé comme donneur pour la valeur manquante y_j et définissons $\mathbf{d} = \{d_{ij}; i \in A_R, j \in A_M\}$. La loi de \mathbf{d} est appelée *mécanisme d'imputation*. Soit w_{ij}^* le facteur appliqué au poids original de l'élément j quand y_i est utilisé pour cet élément. Pour l'élément j , $j \in A_M$,

$$(3) \quad Y_j = \sum_{i \in A_R} w_{ij}^* y_i$$

est la moyenne pondérée des valeurs pour les répondants. Le facteur w_{ij}^* est appelé *fraction d'imputation*, c'est-à-dire la fraction de la réponse manquante y_j que fournit le donneur i . Notons que $w_{ii}^* = 1$ pour $i \in A_R$ et $w_{ij}^* = 0$ pour $i \neq j, i, j \in A_R$. La somme des facteurs d'imputation pour une réponse manquante est contrainte d'être égale à 1,

(4)
$$\sum_{i \in A_R} w_{ij}^* = 1, \quad \forall j \in A.$$

Un estimateur ayant les valeurs imputées définies par (3) et un facteur $w_{ij}^* < 1$ est appelé *estimateur par imputation fractionnaire*.

Nous pouvons écrire un estimateur par imputation linéaire en y sous la forme

$$(5) \quad \theta_I = \sum_{i \in A_R} \sum_{j \in A} w_j w_{ij}^* y_i$$

$$(6) \quad =: \sum_{i \in A_R} \alpha_i y_i,$$

où la notation $A =: B$ signifie que la définition de B est telle qu'il soit égal à A . La somme des $w_j w_{ij}^*$ sur l'ensemble des receveurs pour lesquels i est un donneur (y compris pour lui-même), noté α_i , est le poids total appliqué au donneur i . Si une unité répondante i n'est pas utilisée comme donneur, sauf pour elle-même, alors $\alpha_i = w_i$.

3. Imputation fractionnaire entièrement efficace

Supposons que tous les éléments d'une cellule d'imputation aient la même probabilité de répondre et supposons que les réponses soient indépendantes. Alors, nous pouvons obtenir la loi globale d'un estimateur imputé sous le modèle de réponse en utilisant la structure de probabilité de l'échantillonnage à plusieurs phases, où le modèle de réponse est traité comme étant la deuxième phase du mécanisme d'échantillonnage.

Si les probabilités de réponse dans une cellule sont uniformes, alors un estimateur raisonnable du total est la somme pondérée des estimateurs par le ratio

$$(7) \quad \theta_{FE} = \sum_{g=1}^G \left(\sum_{i \in A_R \cap U_g} w_i \right) \left(\sum_{i \in A_R \cap U_g} w_i y_i \right) / \sum_{i \in A_R \cap U_g} w_i y_i.$$

Dans le contexte de l'échantillonnage à deux phases, Kott et Stukel (1997) ont donné à l'estimateur (7) le nom d'estimateur avec facteur d'extension pondéré. L'estimateur (7) est dit entièrement efficace parce qu'il ne contient aucune variabilité due à la sélection aléatoire des donneurs. Si les w_i sont les mêmes pour tous les éléments d'une cellule, le ratio

$$(8) \quad \left(\sum_{i \in A_R \cap U_g} w_i \right) \left(\sum_{i \in A_R \cap U_g} w_i y_i \right) / \sum_{i \in A_R \cap U_g} w_i y_i$$

est une moyenne simple et, donc, sans biais pour la moyenne de cellule, sachant qu'il existe au moins un répondant dans la cellule. Si les w_i d'une cellule ne sont pas égaux, alors (8) présente un biais de ratio. Il est possible que le nombre d'éléments dans une cellule, n_g , soit positif et

Imputation hot deck pour le modèle de réponse

Wayne A. Fuller et Jae Kwang Kim¹

Résumé

L'imputation hot deck est une procédure qui consiste à remplacer les réponses manquantes à certaines questions par des valeurs empruntées à d'autres répondants. L'un des modèles sur lesquels elle s'appuie est celui où l'on suppose que les probabilités de réponse sont égales dans les cellules d'imputation. Nous décrivons une version efficace de l'imputation hot deck pour le modèle de réponse dans les cellules et donnons un estimateur de la variance dont le traitement informatique est efficace. Nous détaillons une approximation de la procédure entièrement efficace dans laquelle un petit nombre de valeurs sont imputées pour chaque non-répondant. Nous illustrons les procédures d'estimation de la variance dans une étude de Monte Carlo.

Mots clés : Non-réponse, imputation fractionnaire; probabilité de réponse; estimation de la variance par rééchantillonnage.

1. Introduction

Dans les enquêtes par sondage, l'imputation est utilisée comme méthode de traitement de la non-réponse partielle. Dans le cas de l'imputation hot deck, les valeurs imputées sont des fonctions des répondants compris dans l'échantillon courant. Sande (1983) et Ford (1983) décrivent l'imputation hot deck. Kalton et Kasprzyk (1986), ainsi que Little et Rubin (2002) passent en revue diverses procédures d'imputation.

Dans l'une des versions de l'imputation hot deck, la valeur imputée est celle donnée par un répondant appartenant à la même cellule d'imputation, où les cellules d'imputation forment une subdivision exhaustive et disjointe de la population. Dans le cas de l'imputation hot deck aléatoire, des valeurs provenant de répondants appartenant à la même cellule d'imputation sont attribuées au hasard aux non-répondants. L'enregistrement qui fournit la valeur est appelé le *donneur* et celui dans lequel la valeur manque est appelé le *receveur*.

La variance est généralement plus grande pour l'estimateur imputé que pour l'échantillon complet, parce que la non-réponse réduit la taille de l'échantillon et que l'estimateur imputé peut contenir une composante due à l'imputation aléatoire. Rao et Shao (1992) ont proposé pour l'imputation hot-deck une méthode du jackknife ajusté où les unités de la première phase sont sélectionnées avec remise. Rao et Sitter (1995) discutent de la méthode d'estimation de la variance par le jackknife ajusté pour l'imputation par le ratio. Rao (1996) et Sitter (1997) utilisent la méthode du jackknife ajusté dans le cas de l'imputation par la régression. Shao, Chen et Chen (1998) appliquent la notion de Rao et Shao (1992) à la méthode des répliques

Dans le présent article, nous examinons l'imputation hot deck pour une population subdivisée en cellules d'imputation. À la section 2, nous décrivons le modèle de réponse. À la section 3, nous introduisons l'imputation fractionnée entièrement efficace et présentons une méthode d'estimation de la variance pour l'estimateur par imputation, sous l'hypothèse que la probabilité de non-réponse est constante dans une cellule. À la section 4, nous proposons une modification de la méthode entièrement efficace avec utilisation d'un plus petit nombre de donneurs. À la section 6, nous exposons les résultats d'une étude en simulation. Enfin, à la dernière section, nous résumons l'étude.

1. Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA, 50011, États-Unis; Jae Kwang Kim, Department of Applied Statistics, Konkuk University, Seoul, 120-749, Corée.

- Thompson, S.K., et Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.
- Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.
- Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Métron*, 2, 461-493, 646-683.
- Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the U.S. Census Bureau. *Journal of Official Statistics*, 14, 119-135.
- Wang, N., et Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.
- Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.
- Wu, C., et Rao, J.N.K. (2004). Empirical likelihood ratio confidence intervals for complex surveys. Soumis pour publication.
- Wu, C., et Rao, J.N.K. (2005). Article présenté à la réunion 2005 International Statistical Institute, Sydney, Australie.
- Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Griffin.
- Zarkovic, S.S. (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society, Series A*, 119, 336-338.

- Scott, A., et Davis, P. (2001). Estimating interviewer effects for survey responses. *Proceedings of Statistics Canada Symposium 2001*.
- Shao, J. (2002). Resampling methods for variance estimation in complex surveys with a complex design. Dans *Survey Non-response* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc., 303-314.
- Shao, J., et Tu, D. (1995). *The jackknife and the Bootstrap*. New York: Springer Verlag.
- Singh, A.C., et Wu, S. (2001). Estimation for multistage complex surveys by modified regression. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 69-77.
- Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : Problèmes et solutions. *Techniques d'enquête*, 20, 3-15.
- Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- Sitter, R.R., et Wu, C. (2001). A note on Woodruff confidence interval for quantiles. *Statistics & Probability Letters*, 55, 353-358.
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Skinner, C.J., Holt, D. et Smith, T.M.F. (Eds.). (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.
- Smith, T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society, Series A*, 139, 183-204.
- Smith, T.M.F. (1994). Sample surveys 1975-1990: an age of reconciliation? *Revue Internationale de Statistique*, 62, 5-34.
- Sielman, S.V., et Overton, W.S. (1994). Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.
- Sukhrame, P.V. (1947). The problem of plot size in large-scale yield surveys. *Journal of the American Statistical Association*, 42, 297-310.
- Sukhrame, P.V. (1954). *Sampling Theory of Surveys*, with Applications. Ames: Iowa State College Press.
- Sukhrame, P.V., et Panse, V.G. (1951). Crop surveys in India – II. *Journal of the Indian Society of Agricultural Statistics*, 3, 97-168.
- Sukhrame, P.V., et Seth, G.R. (1952). Non-sampling errors in surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.
- Thomas, D.R., et Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.
- Schabab, W.T. (Ed.) (1996). Indirect Estimation in U.S. Federal Programs. New York: Springer
- Schabenberger, O., et Gregory, T.C. (1994). Solutions de remplacement pour les plans *mpi* authentiques : Une étude comparative. *Techniques d'enquête*, 20, 193-200.
- Sämdal, C.-E., Swenson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Sämdal, C.-E., Swenson, B. et Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- Sämdal, C.-E. (1996). Efficient estimators with variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- Salati, M., et Seber, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacements. *Biometrika*, 84, 209-219.
- Statistics, (à paraître).
- Rubin-Bleuer, S., et Schiopu-Kratna, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, (à paraître).
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- Royal, R.M., et Herson, J.H. (1973). Robust estimation in finite populations, I et II. *Journal of the American Statistical Association*, 68, 880-889 et 890-893.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- Royal, R.M., et Cumberland, W.G. (1981). An empirical study of the ratio estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.
- Royal, R.M., et Cumberland, W.G. (1981). An empirical study of the certain linear regression models. *Biometrika*, 57, 377-387.
- Royal, R.M. (1970). On finite population sampling theory under ratio estimators and estimates of its variance. *Journal of the American Statistical Association*, 65, 1269-1279.
- Royal, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.
- Rapport technique, Bureau de la statistique de Suède.
- Rosén, B. (1991). Variance estimation for systematic pps-sampling. *Biometrika*, 74, 1-12.
- Roberts, G., Rao, J.N.K. et Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- Rivera, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214.
- Rivera, L.-P. (2002). Estimating equations for the analysis of survey data using poststratification information. *Samhva*, Series A, 64, 364-378.
- Renssen, R.H., et Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-375.
- Rao, J.N.K., Yung, W. et Hidiroglou, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Samhva*, Series A, 64, 364-378.
- Rao, J.N.K., Jocklyn, W. et Hidiroglou, M.A. (2003). Confidence interval coverage properties for regression estimators in unit-phase and two-phase sampling. *Journal of Official Statistics*, 19.
- Rao, J.N.K., Scott, A.J. et Benham, E. (2003). Défaite les structures des données d'enquête complexes : Théorie élémentaire et applications de l'échantillonnage inverse. *Techniques d'enquête*, 29, 119-131.
- Rao, J.N.K., Hartley, H.O. et Cochran, W.G. (1967). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-491.
- Rao : Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage

- Hansen, M.H., Hurwitz, W.N., Marks, E.S. et Malin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190
- Hansen, M.H., Hurwitz, W.N., Nisselson, H. et Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.
- Hartley, H.O. (1959). Analytical studies of survey data. In Volume in Honour of Corrado Gini, Istituto di Statistica, Rome, 1-32.
- Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sanhya*, Series C, 36, 99-118.
- Hartley, H.O., et Biemer, P. (1978). The estimation of nonsampling variances in current surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 257-262.
- Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.
- Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- Hartley, H.O., et Rao, J.N.K. (1978). The estimation of nonsampling variance components in sample surveys. Dans *Survey Measurement* (Ed. N.K. Namoodini), New York: Academic Press, 35-43.
- Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa, Etats-Unis.
- Hintkins, S., Oh, H.T. et Scheuren, F. (1997). Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 13-24.
- Holl, D., et Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- Hubback, J.A. (1927). Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (représenté dans *Sanhya*, 1946, vol. 7, 281-294).
- Hussain, M. (1969). Construction of regression weights for estimation in sample surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.
- Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.
- Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, 129-154.
- Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.
- Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, A13, 1919-1939.
- Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changing in the probabilities. *Journal of the American Statistical Association*, 46, 105-109.
- Statistique Canada, N° 12-001-XPB au catalogue
- Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo, Central Bureau of Statistics of Norway.
- Kim, J., et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.
- Kim, J.K., Brick, J.M., Fuller, W.A. et Kalton, G. (2004). On the bias of the multiple imputation variance estimator in survey sampling. Rapport technique.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L. (1995). The hundred year's wars of survey sampling. *Statistics in Transition*, 2, 813-830.
- Kish, L., et Scott, A.J. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- Kish, L., et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 384-389.
- Kott, P.S. (2005). Randomized-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263-277.
- Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- Kruskal, W.H., et Mosteller, F. (1980). Representative sampling IV: The history of the concept in Statistics, 1895-1939. *Revue Internationale de Statistique*, 48, 169-195.
- Laplace, P.S. (1820). A philosophical essay on probabilities. English translation, Dover, 1951.
- Lavallée, P. (2002). *Le Sondage indirect, ou la Méthode généralisée du parage des poids*. Éditions de l'Université de Burkelles, Belgique, Éditions Ellipse, France.
- Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
- Lehtonen, R., et Pukkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.
- Lindley, D.V. (1996). Letter to the editor. *American Statistician*, 50, 197.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury.
- Lohr, S.L., et Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 2710280.
- Lohr, S.L., et Rao, J.N.K. (2005). Multiple frame surveys: point estimation and inference. *Journal of the American Statistical Association* (en révision).
- Lu, W.W., Brick, M. et Sitter, R.R. (2004). Algorithms for constructing combined strata grouped jackknife and balanced repeated replication with domains. Rapport technique, Westat, Rockville, Maryland.
- Mach, L., Reiss, P.T. et Schiopu-Kratina, I. (2005). The use of the transportation problem in co-ordinating the selection of samples for business surveys. Rapport technique HSMJD-2005-006E, Statistique Canada, Ottawa.

- Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.
- Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- Cochran, W.G. (1940). The estimation of the yield of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science*, 30, 262-275.
- Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 191-212.
- Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples from a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- Cochran, W.G. (1953). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Dalenius, T. (1957). *Sampling in Sweden*. Stockholm: Almqvist and Wiksell.
- Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- Deming, W.E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.
- Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.
- Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *The Annals of Mathematical Statistics*, 11, 427-444.
- Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.
- Deville, J., et Särndal, C.-E. (1992). Calibration estimators in survey. *Journal of the American Statistical Association*, 87, 376-382.
- Durbin, J. (1968). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute*, 36, No. 3, 113-119.
- Efron, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.
- Ernst, L.R. (1989). Weighing issues for longitudinal household and family estimates. Dans *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons, Inc., 135-169.
- Ernst, L.R. (1999). The maximization and minimization of sample overlap problem: A half century of results. *Bulletin of the International Statistical Institute*, Vol. LVII, Book 2, 293-296.
- Far, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- Felllegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Felllegi, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS sampling without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington DC, 434-442.
- Felllegi, I.P. (1981). Should the census counts be adjusted for allocation purposes? - Equity considerations. Dans *Current Topics in Survey Sampling* (Eds. D. Krewski, R. Plehak et J.N.K. Rao). New York: Academic Press, 47-76.
- Frankis, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā, Series C*, 37, 117-132.
- Fuller, W.A. (1995). Estimation in the presence of measurement error. *Revue Internationale de Statistique*, 63, 121-147.
- Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331-345.
- Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- Fuller, W.A., et Burneister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.
- Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Technique d'enquête*, 27, 49-56.
- Gambino, J., Kennedy, B. et Singh, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 269-278.
- Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 310-328.
- Godambe, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 28, 376-382.
- Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- Graubard, B.I., et Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.
- Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.
- Hájek, J. (1971). Comments on a paper by Basu, D. Dans *Foundations of Statistical Inference* (Eds. V.P. Godambe et D.A. Sprott). Toronto: Holt, Rinehart and Winston.
- Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- Hansen, M.H., Dalenius, T. et Tepping, B.J. (1985). The development of sample surveys of finite populations. Chapter 13 in *A Celebration of Statistics. The ISI Centenary Volume*. Berlin: Springer-Verlag.
- Hansen, M.H., Hurwitz, W.N. et Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 339-374.
- Hansen, M.H., Hurwitz, W.N. et Meadow, W.G. (1953). *Sample Survey Methods and Theory*. Vols. I et II. New York: John Wiley & Sons, Inc.
- Hansen, M.H., Meadow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

obtenus sont simples, mais pas nécessairement efficaces par rapport aux estimateurs optimaux de Hájek (1974) ou aux estimateurs du PMV. La méthode MGPP a une grande applicabilité et mérite l'attention des praticiens.

9. Conclusion

L'apport de Joe Waksberg à la théorie et aux méthodes des enquêtes par sondages reflète bien l'interaction entre la théorie et la pratique. Dans le cadre de son travail au Censur Bureau des États-Unis, puis à Westat, il a fait face à de réels problèmes d'ordre pratique et a souvent trouvé des solutions théoriques judicieuses. Par exemple, dans un article marquant (Waksberg 1978), il a décrit une ingénieuse méthode de composition aléatoire (CA) qui réduit considérablement les coûts d'enquête par rapport à la composition de numéros entièrement au hasard. Il a présenté des arguments théoriques solides pour en démontrer l'efficacité. L'utilisation généralisée des enquêtes par CA est due pour une bonne part à l'argumentation théorique de Waksberg (1978) et à des perfectionnements ultérieurs. Joe Waksberg est un spécialiste de l'échantillonnage d'enquête que j'admire énormément et je suis très honoré d'avoir reçu le prix Waksberg 2005 pour les techniques d'enquête.

Remerciements

Je tiens à remercier David Bellhouse, Wayne Fuller, Jack Gambino, Graham Kalton, Fritz Scheuren et Sharon Lohr, dont les observations et les suggestions m'ont été très utiles.

Bibliographie

Aïres, N., et Rosen, B. (2005). On inclusion probabilities and relative estimator bias for Pareto rps sampling. *Journal of Statistical Planning and Inference*, 128, 543-567.

Andreatta, G., et Kaufmann, G.M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81, 657-666.

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bankier, M.D. (2003). 2001 Canadian Census weighting: switch from projection GREG to pseudo-optimal regression estimation. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Rapport technique no. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

Bankier, M.D., Rathwell, S. et Matkowsky, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, direction de la méthodologie, division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. Dans *Foundations of Statistical Inference* (Eds. V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Bellhouse, D.R., et Rao, J.N.K. (2002). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, 102, 47-58.

Bellhouse, D.R., et Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

Bellhouse, D.R., et Stafford, J.E. (2001). Régression polynomiale locale dans le cas des enquêtes complexes. *Techniques d'enquête*, 27, 219-226.

Bellhouse, D.R., Chipman, H.A. et Stafford, J.E. (2004). Additive models for survey data via penalized least squares. Rapport technique.

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue internationale de Statistique*, 51, 279-292.

Binder, D.A., et Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.

Binder, D.A., Kovacevic, M. et Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. I, 6-62.

Brackstone, G. (2002). Stratégies et approches relatives aux statistiques régionales. *Techniques d'enquête*, 28, 125-133.

Brackstone, G., et Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sanhyā, Series C*, 42, 97-114.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Brewer, K.R.W., et Hanif, M. (1983). *Sampling With Unequal Probabilities*. New York: Springer-Verlag.

Buskirk, T.D., et Lohr, S.L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Casady, R.J., et Valliant, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L., et Skinner, C.J. (Eds.) (2003). *Analysis of Survey Data*. Chichester: Wiley.

Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 12, 1223-1239.

Chen, J., Chen, S.Y., et Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53-68.

d'imputation marginale comme celles du plus proche voisin et du donneur aléatoire, on traite souvent des classes d'imputation. Malheureusement, on traite souvent les valeurs imputées comme si l'on s'agissait de valeurs vraies, puis on calcule des estimations et des estimations de la variance. Les estimations ponctuelles imputées de paramètres marginaux sont généralement valides en présence d'un marquant, Hartley (1974) a calculé des estimateurs à double base « optimaux » pour des plans d'échantillonnage généraux et des unités d'observation pouvant être différentes dans les deux bases. Fuller et Burmeister (1972) ont proposé des estimateurs « optimaux » améliorés. Toutefois, les estimateurs optimaux utilisent des ensembles de poids différents pour chaque élément y , ce qui n'est pas souhaitable dans la pratique. Skinner et Rao (1996) ont calculé, pour les enquêtes à double base, des estimateurs du pseudo-maximum de vraisemblance (PMV) qui utilisent le même ensemble de poids pour tous les éléments y , comme dans le cas des estimateurs « à base unique » (Kallott et Anderson 1986), et qui maintiennent l'efficacité. Lohr et Rao (2005) ont formulé une théorie unifiée des conditions des enquêtes à bases multiples en prolongeant les estimateurs optimaux, PMV et à base unique. Lohr et Rao (2000, 2005) ont obtenu des estimateurs de variance jackknife asymptotiquement valides. Ces résultats généraux méritent l'attention des praticiens lorsqu'on travaille avec deux ou plusieurs bases. Les enquêtes téléphoniques à double base (téléphones cellulaires et téléphones fixes) nécessitent l'attention des théoriciens, car on ignore comment pondérer dans le cas de partages un téléphone cellulaire, d'autres en possèdent un pour chaque personne.

8.6 Échantillonnage indirect

On peut utiliser la méthode de l'échantillonnage indirect lorsqu'on ne dispose pas de la base d'une population cible U_B mais qu'on emploie la base d'une autre population U_A , liée à U_B , pour tirer un échantillon probabiliste. On utilise les liens entre les deux populations pour établir des poids appropriés qui peuvent donner des estimateurs sans biais et des estimateurs de variance. Lavallée (2002) a mis au point une méthode unifiée, appelée méthode généralisée du partage des poids (MGPP), inspirée de plusieurs méthodes connues : la méthode du partage des poids d'Ernst (1989) pour l'estimation transversale à partir d'enquêtes-ménages longitudinales, l'échantillonnage par réseau et l'estimation de la multiplicité (Stiksen 1970), ainsi que l'échantillonnage en grappes adapté (Thompson et Seber 1996). La théorie de Rao (1968) sur l'échantillonnage à partir d'une base contenant une quantité inconnue de doubles comptes peut être considérée comme un cas particulier de la MGPP. On peut aussi employer la MGPP pour travailler avec des bases multiples, les estimateurs ainsi

8.5 Enquêtes à bases multiples

Les enquêtes à bases multiples emploient deux ou plusieurs bases chevauchantes pour couvrir entièrement la population cible. Hartley (1962) a étudié le cas particulier d'une base complète B , d'une base incomplète A et d'un échantillonnage aléatoire simple mené indépendamment dans les deux bases. Il a montré que par rapport à l'estimateur à base unique complète, un estimateur à double base « optimal » pouvait donner lieu à d'importants gains d'efficacité pour le même coût, à condition que le coût par unité pour la base A soit nettement inférieur au coût par

tiennent l'attention des praticiens.

méthodes d'estimation de la variance susmentionnées méthodes d'estimation de la variance au hasard. Les uniques utilisant une seule valeur imputée au hasard. Les variances due à l'imputation par rapport à l'imputation (1996). L'imputation partielle offre l'avantage de réduire la quement valides; voir aussi Kallott et Kish (1984) et Fay méthode donnait également des inférences asymptotiques plus d'une valeur imputée au hasard et ont montré que cette Fuller (2004) ont étudié l'imputation partielle en utilisant de la variance au moyen de l'imputation simple. Kim et (2002) et Rao (2000, 2005) sur les méthodes d'estimation fois. Le lecteur est invité à consulter les articles de Shao valides à partir d'ensembles de données imputées une seule réalisant des inférences efficaces et asymptotiquement dernières années, on a fait des progrès impressionnants en efficacité opérationnelle et de rentabilité. Au cours des on préfère souvent l'imputation simple pour des raisons Brick, Fuller et Kalton (2004) et d'autres auteurs. En outre, (1996), Binder et Sun (1996), Wang et Robins (1998), Kim, certaines difficiles, comme en font état Koit (1995), Fay estimateurs de variance à imputation multiple comportent moyen des imputations multiples. Malheureusement, les moyenne des estimateurs nés de la variance obtenus au somme des cartes entre estimateurs imputés est ajoutée à la imputation multiple peut régler ce problème parce qu'une de Rubin (1987) soutiennent que l'estimateur de variance à valeurs manquantes. Les partisans de l'imputation multiple compte la variabilité supplémentaire due à l'estimation des de la variance de l'estimateur imputé, faute de prendre en grands échantillons, notamment une forte sous-estimation peuvent produire des inférences erronées, même pour de théorique. Mais les estimateurs « nés » de la variance mécanisme de réponse ou d'un modèle d'imputation hypomarginaux sont généralement valides en présence d'un marquant, Hartley (1974) a calculé des estimateurs à double listes incomplètes contenant de fortes proportions de sans-abri et de personnes atteintes du SIDA, lorsque des conventions partiellement à l'échantillonnage de population rares ou difficiles à joindre, comme les populations de

Rao : Evaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage

qui concerne les graphiques, l'estimation, les tests et la sélection de paramètres « de lissage » pour ajuster les modèles.

8.3 Erreurs de mesure

Habituellement, on suppose que les erreurs de mesure sont additives et que leur moyenne est nulle. Par consé-

sont additives et que leur moyenne est nulle. Par conséquent, les estimateurs habituels du total et des moyennes résistent sans biais no no no convergents. Toutefois, cette caractéristique positive n'est pas nécessairement valable pour des paramètres plus complexes comme la fonction de distribution, les quantiles et les coefficients de régression. Dans ce dernier cas, les estimateurs habituels sont biaisés, même pour de grands échantillons, et peuvent donc produire des inférences erronées (Fuller 1995). Il est possible d'obtenir des estimateurs corrigés pour le biais si l'on dispose d'estimations des variances de l'erreur de mesure. On peut obtenir ces dernières en affectant des ressources, à l'étape de l'élaboration du plan de sondage, pour faire des observations répétées sur un sous-échantillon. Fuller (1975, 1995) a préconisé l'utilisation de méthodes appropriées en présence d'erreurs de mesures, et les méthodes corrigées pour le biais méritent l'attention des praticiens.

mentent l'attention des praticiens. Hartley et Rao (1978) et Hartley et Biemer (1978) ont établi des conditions d'affectation des intervieweurs et des codeurs qui permettent d'estimer les variances d'échantillonage et de réponse pour la moyenne arithmétique ou le total à partir d'enquêtes courantes. Malheureusement, le plan de sondage des enquêtes d'aujourd'hui satisfait rarement ces conditions et, même si c'était le cas, on dispose rarement de l'information requise sur les affectations des intervieweurs et des codeurs à l'étape de l'estimation.

On utilise souvent les composantes linéaires des modèles de variance pour estimer la variabilité des intervieweurs. Ces modèles sont appropriés pour la réponse continue, mais pas pour les réponses binaires. L'approche du modèle linéaire pour les réponses binaires peut entraîner une sous-estimation des corrélations intra-intervieweurs. Scott et Davis (2001) ont proposé des modèles hiérarchiques pour les réponses binaires afin d'estimer la variabilité due aux intervieweurs. Comme les réponses sont souvent binaires dans bon nombre d'enquêtes, les praticiens doivent prêter attention à ces modèles pour effectuer des analyses pertinentes des données d'enquête avec réponses binaires.

8.4 Imputation des données d'enquête manquantes

Dans la pratique, on utilise couramment l'imputation pour remplacer des éléments manquants. On s'assure ainsi que les résultats d'analyses différentes de l'ensemble de données complètes sont cohérents entre eux en utilisant le même poids d'échantillonnage pour tous les éléments. Bon nombre d'organismes statistiques utilisent des méthodes

Également une approche systématique de l'estimation par calage et de l'intégration des enquêtes. Le lecteur est invité à consulter les articles de Rao (2004) et Wu et Rao (2005). Il reste encore à perfectionner ces notions, notamment en ce qui concerne la pseudo-vraisemblance empirique, mais la théorie de la VE dans le contexte des enquêtes mérite l'attention des praticiens.

8.2 Analyses exploratoires des données d'enquête

Dans la section 6, nous avons abordé les méthodes d'analyse confirmative de données d'enquête tenant compte du plan de sondage, comme l'estimation ponctuelle des paramètres de modélisation (ou de recensement) et des erreurs-types associées, ainsi que les tests formels d'hypothèses. Les graphiques et les analyses exploratoires des données d'enquête sont aussi très utiles. Ces méthodes ont fait l'objet d'une foule d'études dans la documentation courante. Encore récemment, certains ajouts à ces méthodes modernes ont été signalés dans la documentation sur les enquêtes et ils méritent l'attention des praticiens. J'en aborde brièvement un certain nombre. Premièrement, on utilise couramment des estimations non paramétriques de densité du noyau pour présenter la forme d'un ensemble de données sans recourir à des modèles paramétriques. On peut aussi les utiliser pour comparer différentes sous-populations.

Bellhouse et Starfford (1999) ont proposé des estimateurs de densité du noyau qui tiennent compte du plan d'échantillonnage en ont étudié les propriétés et ont appliqué les méthodes aux données de l'Enquête sur la santé en Ontario, Buskirk et Lohr (2005) ont étudié les propriétés asymptotiques et les propriétés de population finie des estimateurs de densité du noyau et ont obtenu des bandes de confiance. Ils ont appliqué les méthodes aux données de deux enquêtes américaines, la National Crime Victimization Survey et la National Health and Nutrition Examination Survey.

Deuxièmement, Bellhouse et Starfford (2001) ont mis au point des méthodes de régression polynomiale locale qui

Ontario. Cette approche offre de nombreux avantages en ce

8. Certains aspects théoriques méritant l'attention des praticiens et vice versa

Dans la présente section, j'aborde brièvement quelques exemples d'aspects théoriques importants qui existent mais qui sont peu utilisés dans la pratique.

8.1 Inférence par la vraisemblance empirique

La théorie traditionnelle de l'échantillonnage portait dans une large mesure sur l'estimation ponctuelle et les erreurs-types associées, faisant appel à des approximations normales pour déterminer des intervalles de confiance à l'égard des paramètres d'intérêt. En statistique courante, l'approche de la vraisemblance empirique (VE) (Owen 1988) a beaucoup attiré l'attention en raison de plusieurs propriétés souhaitables. Elle offre une vraisemblance non paramétrique, ce qui donne des intervalles de confiance de VE semblables aux intervalles de vraisemblance paramétrique. La forme et l'orientation des intervalles de VE sont entièrement déterminées par les données; les intervalles pré-servent l'étendue tout en respectant la transformation et, contrairement aux intervalles symétriques de la théorie normale, ils sont particulièrement utiles puisqu'ils donnent des taux d'erreur équilibrés de la queue. Comme je l'ai mentionné dans la section 3.1, Hartley et Rao (1968) ont été les premiers à proposer l'approche de la VE dans le contexte des enquêtes par sondage, mais leur démarche était axée sur des questions d'inférence liées à l'estimation ponctuelle. Chen, Chen et Rao (2003) ont obtenu des intervalles de VE sur la moyenne de population sous échantillonnage aléatoire simple et sous échantillonnage aléatoire stratifié pour des populations contenant bien des zéros. On trouve ces populations dans le contrôle par sondage, où y dénote le montant d'argent dû à l'État et la moyenne arithmétique \bar{y} correspond au montant moyen des créances excessives. Des travaux antérieurs sur le contrôle par sondage ont utilisé des intervalles de vraisemblance paramétrique fondés sur des distributions de mélanges paramétriques pour la variable y . Ces intervalles donnent de meilleurs résultats que les intervalles-types de la théorie normale, mais les intervalles de VE donnent de meilleurs résultats en présence d'écarts par rapport au modèle hypothétique de mélanges, en donnant un taux de non-couverture inférieur à la borne inférieure plus proche du taux d'erreur nominal, ainsi qu'une bonne inférieure plus grande. Pour les plans généraux, Wu et Rao (2004) ont utilisé une pseudo-vraisemblance empirique (Chen et Sitter 1999) pour obtenir des intervalles de pseudo-VE rajustés sur la moyenne arithmétique et la fonction de distribution qui tiennent compte des caractéristiques du plan, et ils ont montré que les intervalles donnaient des taux d'erreur de la queue plus équilibrés que dans le cas des intervalles de la théorie normale. La méthode VE offre

fiables et d'une validation complète des modèles au moyen d'évaluations internes et externes. Bon nombre de méthodes axées sur les effets aléatoires et utilisées dans la théorie statistique courante sont pertinentes à l'estimation sur petits domaines, dont la méthode du meilleur prédicteur empirique (ou méthode de Bayes), celle du meilleur prédicteur linéaire sans biais empirique et celle du modèle hiérarchique bayésien fondé sur des lois de distribution a priori des paramètres de modélisation. Rao (2003) donne une description complète de ces méthodes. La pertinence (sur le plan pratique) et l'intérêt (sur le plan théorique) de l'estimation sur petits domaines ont attiré l'attention de nombreux chercheurs, d'où la réalisation de progrès importants dans l'estimation ponctuelle et celle de l'erreur quadratique moyenne. Dans le monde entier, les « nouvelles » méthodes ont été appliquées avec succès à divers problèmes liés aux petits domaines. Aux États-Unis, on a utilisé récemment des méthodes fondées sur un modèle pour produire des estimations par comté et par district scolaire relativement aux enfants pauvres d'âge scolaire. Chaque année, le département de l'Éducation des États-Unis accorde aux comtés des fonds de plus de sept milliards de dollars sur la base d'estimations par comté fondées sur un modèle. Les fonds alloués soutiennent des programmes d'éducation compensatoire pour répondre aux besoins des enfants défavorisés sur le plan scolaire. Le lecteur trouvera dans Rao (2003, exemple 7.1.2) des renseignements sur cette application. Au Royaume-Uni, le Office of National Statistics a mis sur pied un projet d'estimation sur petits domaines pour établir des estimations fondées sur un modèle au niveau des sections électorales (quelque 2 000 ménages). Schabiel (1996) décrit la pratique et les méthodes d'estimation des programmes statistiques fédéraux des États-Unis qui utilisent des estimateurs indirects pour produire des estimations publiées. Singh, Gambino et Mantel (1994) et Brackstone (2002) traitent de certains aspects pratiques et stratégiques de la statistique des petits domaines. L'estimation sur petits domaines constitue un exemple frappant de l'interaction entre la théorie et la pratique. Les nombreux de questions d'ordre pratique nécessitent une plus grande attention de la part des théoriciens, notamment les suivantes : a) des estimateurs d'échantillage fondés sur un modèle pour concorder avec des estimateurs directs fiables au niveau des grands domaines; b) l'établissement et la validation de modèles de liaisons appropriés et l'étude de questions comme les erreurs dans les variables, la spécification incorrecte du modèle de liaison et les variables omises; c) la mise au point de méthodes qui satisfont plusieurs objectifs : de bonnes estimations spécifiques au domaine, de bons rangs et un bon histogramme des petits domaines.

conséquent, les estimateurs fondés sur un modèle peuvent être fortement biaisés et les inférences risquent d'être erronées. Pfeffermann et ses collègues ont proposé une nouvelle approche de l'inférence sous échantillonnage informatif (voir Pfeffermann et Sverchkov 2003), qui semble donner des inférences plus efficaces que l'approche pondérée et mérité certainement l'attention des utilisateurs de données d'enquête. Toutefois, il reste beaucoup de travail à accomplir, surtout en ce qui concerne le traitement de données fondées sur l'échantillonnage à plusieurs degrés.

Skinner, Holt et Smith (1989), Chambers et Skinner (2003) et Lehtonen et Pakkinen (2004) donnent d'excellentes descriptions des méthodes d'analyse de données d'enquête complexes.

7. Estimation sur petits domaines

Dans les sections précédentes, nous avons abordé les méthodes traditionnelles qui utilisent des estimateurs directs par domaine fondés sur des observations d'échantillons spécifiques aux domaines et sur des données auxiliaires sur la population. Toutefois, ces méthodes ne donnent pas nécessairement des inférences fiables lorsque la taille des échantillons du domaine est infime, voire nulle pour certains domaines. Dans la documentation, les données ou sous-populations dont la taille est infime ou nulle sont appelées petits domaines. Au cours des dernières années, la demande de statistiques fiables sur les petits domaines a grandement augmenté en raison du recours croissant à la statistique des petits domaines dans la formulation des politiques et des programmes. Manifestement, il est rarement possible d'obtenir des échantillons dont la taille globale est assez grande pour soutenir des estimations directes fiables pour tous les domaines d'intérêt. De plus, dans la pratique, il n'est pas possible de prévoir toutes les utilisations des données d'enquête et « le client exige toujours plus qu'il n'est spécifié à l'étape de l'élaboration du plan de sondage » (Füller 1999, page 344). Pour faire des estimations sur petits domaines avec un niveau suffisant de précision, il faut souvent utiliser des estimateurs « indirects » qui empruntent de l'information à des domaines connexes par le biais de données auxiliaires, comme celles du recensement et les données administratives courantes, pour accroître la taille « effective » des échantillons à l'intérieur des petits domaines.

Aujourd'hui, on s'entend à reconnaître que des données explicites liant les petits domaines par le biais de données auxiliaires et venant compenser de la variation résiduelle entre domaines par le biais des effets aléatoires des petits domaines sont nécessaires pour calculer des estimateurs indirects. Le succès des méthodes fondées sur un modèle dépend fortement de la disponibilité de données auxiliaires

Dans certaines conditions, on peut faire abstraction de la situation d'échantillonnage à deux phases où le recensement est l'échantillon de première phase tiré de la superpopulation et l'échantillon est un échantillon probabiliste tiré de la population de recensement. Récemment, on a mené des travaux utiles sur l'estimation de la variance à deux phases lorsque les paramètres de modélisation sont les paramètres cibles (Graubard et Korn 2002; Rubin-Bjeuer et Schiopu Kratna 2005), mais il faudrait approfondir ces travaux pour surmonter la difficulté de spécifier la structure de covariance présente une difficulté : la solution $\theta^{(r)}$ n'existe pas nécessairement pour certaines répétitions bootstrap r (Binder, Kovacevic et Roberts 2004). Rao et Tausi (2004) ont utilisé la méthode du bootstrap avec fonction d'estimation, qui évite la difficulté. Selon cette méthode, on résout $U(\theta) = U^{(r)}(\theta)$ pour θ en utilisant une seule étape de l'itération de Newton-Raphson avec θ comme valeur de départ. On utilise alors dans (10) l'estimateur $\theta^{(r)}$ ainsi obtenu pour calculer l'estimateur bootstrap avec fonction d'estimation de la variance de θ qu'on peut facilement mettre en œuvre à partir du fichier de données qui fournit les poids de rééchantillonnage, en modifiant légèrement un projeté qui tient compte des poids d'échantillonnage. Il est intéressant de noter que l'estimateur bootstrap avec fonction d'estimation de la variance équivalait à un estimateur de sandwich par linéarisation de Taylor de la variance qui utilise l'estimateur bootstrap de la variance de $U(\theta)$ et l'inverse de la matrice d'information observée (dérivée de $-U(\theta)$), tous deux évalués à $\theta = \theta$ (Binder et coll. 2004).

Contrairement aux méthodes de rééchantillonnage, les méthodes de linéarisation de Taylor produisent des estimateurs de variance asymptotiquement valides pour les plans d'échantillonnage généraux, mais elles nécessitent une formule distincte pour chaque estimateur θ . Binder (1983), Rao, Yung et Hidiroglou (2002) et Demnati et Rao (2004) ont fourni des formules unifiées d'estimation de variance par linéarisation pour des estimateurs définis comme des solutions aux équations d'estimation.

Pfeffermann (1993) a étudié le rôle des poids de sondage dans l'analyse des données d'enquête. Si le modèle de population est valable pour l'échantillon (c'est-à-dire s'il est sans biais d'échantillonnage), les estimateurs non pondérés fondés sur un modèle sont alors plus efficaces que les estimateurs pondérés et donnent des inférences valides, notamment pour des données où la taille des échantillons est faible et la variation des poids est élevée. Toutefois, pour les données ordinaires provenant d'enquêtes à grande échelle, le plan d'enquête est informatif et le modèle de population n'est pas nécessairement valable pour l'échantillon. Par

d'échantillon diminue, ce qui entraîne des taux d'erreur de type I démesurément élevés par rapport aux niveaux nominaux, contrairement aux corrections du deuxième ordre de Rao-Scott (Thomas et Rao 1987). Les corrections du premier et du deuxième ordre sont maintenant appelées corrections de Rao-Scott et constituent des options par défaut dans la nouvelle version du SAS. Roberts, Rao et Kumar (1987) ont mis au point des corrections du type Rao-Scott pour les tests d'analyse de régression logistique des proportions estimatives des cellules associées à une variable de réponse binaire. Ils ont appliqué les méthodes aux moyennes de domaine provenant d'une enquête sur la fécondité menée au Fidji, recoupées par niveau de scolarité et par nombre d'années depuis le premier mariage de la femme, une moyenne de domaine étant le nombre moyen d'enfants nés de femmes de race indienne appartenant au domaine.

Dans le contexte des enquêtes à grande échelle utilisant des plans d'échantillonnage stratifié à plusieurs degrés, les méthodes de rééchantillonnage ont fait l'objet de nombreuses études. Pour les besoins de l'inférence, les UPF de l'échantillon sont traitées comme si elles étaient tirées avec remise à l'intérieur des strates. Les variances s'en trouvent surestimées, mais cette surestimation est faible si la fraction d'échantillonnage globale des UPF est négligeable. Soit θ l'estimateur pondéré d'un paramètre « de recensement » d'intérêt, calculé d'après les poids finaux $w_{i(r)}$, et soient les poids correspondant à chaque pseudo-répétition r produits par la méthode de rééchantillonnage dénotés par $w_{i(r)}$. L'estimateur fondé sur les pseudo-poids de rééchantillonnage $w_{i(r)}$ est dénoté $\hat{\theta}_{(r)}$ pour chaque $r = 1, \dots, R$. Un

alors la forme

$$v(\hat{\theta}) = \sum_{r=1}^R c_r (\hat{\theta}_{(r)} - \hat{\theta})^2 \quad (10)$$

pour les coefficients spécifiés c_r dans (10) déterminés par la méthode de rééchantillonnage.

Les méthodes de rééchantillonnage couramment utilisées comprennent : a) le jackknife avec suppression d'une grappe (ou d'une UPF), b) la répétition compensée (*balanced repeated replicate* ou BRR), notamment pour $n_h = 2$ UPF dans chaque strate h , et c) le bootstrap de Rao et Wu (1988). On obtient les pseudo-répétitions jackknife en supprimant tour à tour chaque grappe d'échantillon $r = (h)_r$, et les poids de sondages jackknife $d_{i(r)}$ prennent la valeur 0 si l'unité d'échantillonnage i est dans la grappe

ponctuelle des paramètres d'intérêt pour calculer les bons estimateurs et les erreurs-types, comme nous le démontrons ci-dessous. Les méthodes d'inférence par rééchantillonnage ont donc attiré l'attention des utilisateurs, qui peuvent très facilement effectuer les analyses eux-mêmes à l'aide de logiciels standard. Toutefois, la mise en circulation de fichiers de données à grande diffusion avec poids de rééchantillonnage risque d'entraîner des problèmes de confidentialité, comme l'identification des grappes à partir des poids de rééchantillonnage. Les théoriciens ont d'ailleurs un défi à relever : celui de mettre au point des méthodes appropriées qui préservent la confidentialité des données. Lu, Brick et Sitter (2004) ont proposé de regrouper les strates et de former des pseudo-répétitions en utilisant les strates combinées pour l'estimation de la variance, limitant ainsi le risque d'identification des grappes à partir du fichier de données à grande diffusion ainsi obtenu. Le groupement des strates ou des UPF à l'intérieur des strates simplifie l'estimation de la variance en réduisant le nombre de pseudo-répétitions utilisés, comparativement à la méthode jackknife avec suppression d'une grappe, qu'on utilise couramment et que nous abordons ci-dessous. Une méthode d'échantillonnage inverse servant à défaire la structure complexe des données d'enquête tout en offrant une protection contre la révélation des étiquettes de grappe (Hinkins, Oh et Schuren 1997; Rao, Scott et Benhin 2003) semble prometteuse, mais il reste beaucoup de travail à accomplir sur les méthodes d'échantillonnage inverse avant qu'elle n'intéresse l'utilisateur.

Rao et Scott (1981, 1984) ont mené une étude systématique de l'effet du plan de sondage sur le test chi carré et le test du rapport des vraisemblances, tests standardisés associés à un tableau multiple de comptes estimatifs ou de proportions. Ils ont montré que la variable à tester était asymptotiquement distribuée sous forme de somme pondérée de variables χ^2 indépendantes, les poids étant les valeurs propres d'une matrice d'« effets généralisés du plan de sondage ». Ce résultat général montre que le plan d'enquête peut avoir un effet important sur le taux d'erreur de type I. Rao et Scott ont proposé des corrections simples du premier ordre aux statistiques chi carré standardisées, qu'on peut calculer à partir de tableaux publiés comprenant des estimations des effets du plan de sondage pour les cellules d'estimation et leurs totaux marginaux, ce qui facilite les analyses secondaires à partir de tableaux publiés. Ils ont également calculé des corrections du deuxième ordre qui sont plus exactes, mais qui nécessitent la connaissance d'une matrice complète des covariances estimatives des cellules d'estimations, comme dans le cas des tests de Wald, bien connus. Toutefois, les tests de Wald peuvent devenir très instables lorsque le nombre de cellules d'un tableau multiple augmente et que le nombre de grappes

d'adopter de nouvelles mesures de la taille en utilisant la méthode de Keyfitz à l'intérieur de chaque groupe aléatoire. Par contre, les stratégies (PIPT, NHT) ne conviennent pas tellement à cette fin (Felleig 1966). Je crois savoir qu'on utilise souvent la stratégie de Rao-Hartley-Cochran en contrôle par sondage et dans d'autres applications comptables. Murthy (1957) a utilisé un plan sans PIPT fondé sur le tirage d'unités successives avec probabilités $p_1, p_2, \dots, p_j, \dots, p_k / (1 - p_1 - p_2 - \dots - p_j)$ et ainsi de suite, et l'estimateur suivant :

$$Y^m = \sum_{i=1}^n y_i^m \frac{d(s_i)}{d(s)}, \quad (9)$$

où $d(s_i)$ est la probabilité conditionnelle d'obtenir l'échantillon s lorsque l'unité i a été sélectionnée en premier. Il a également proposé un estimateur de variance non négatif nécessitant les probabilités conditionnelles, $d(s_i, j)$, d'obtenir s lorsque i et j sont sélectionnés dans les deux premiers tirages. Pendant plusieurs années, les praticiens ont accordé peu d'attention à cette méthode à cause de la complexité des calculs mais, plus récemment, on l'a appliquée dans des domaines inattendus, dont la découverte de pétrole (Andreatta et Kaufmann 1986) et l'échantillonnage séquentiel, dont l'échantillonnage inverse et certains schémas d'échantillonnage adaptable (Salehi et Seber 1997). Il convient de noter qu'au cours des dernières années, on s'est beaucoup intéressé à l'échantillonnage adaptable puisqu'il s'agit d'une méthode d'échantillonnage efficiente pour estimer des totaux ou des moyennes de populations rares (Thompson et Seber 1996). Dans son application à la découverte de pétrole, le schéma d'échantillonnage successif est une caractérisation de la découverte et l'ordre dans lequel les champs de pétrole sont découverts est déterminé par l'échantillonnage proportionnel à la taille des champs et sans remise, selon un vieux principe de l'industrie : « en moyenne, on trouve d'abord les grands champs ». Ici, $p_i = y_i / Y$ et la réserve de pétrole totale Y est présuée connue d'après des critères géologiques. Dans cette application, les géologues s'intéressent à la distribution par taille de tous les champs du bassin et, après l'exploration partielle d'un bassin, l'échantillon est composé de grands y_i de dépôts découverts. On peut estimer la fonction de distribution par taille $F(a)$ en utilisant l'estimateur de Murthy (9) dans lequel y_i est remplacé par la variable indicatrice $I(y_i \leq a)$. Le calcul de $d(s_i)$ et $p(s_i)$, toutefois, est très complexe, même pour des échantillons de taille moyenne. Afin de surmonter cette difficulté de calcul, Andreatta et Kaufmann (1986) ont utilisé des représentations intégrales de ces quantités pour formuler des expressions asymptotiques de l'estimateur de Murthy, dont les premiers termes sont aisés à calculer. De même, ils obtiennent des approximations calculables de l'estimateur de variance de

6. Analyse des données d'enquête et des méthodes de rééchantillonnage

Murthy. Il est à noter qu'on ne peut employer ici l'estimateur NHT de $F(a)$, car les probabilités d'inclusion sont des fonctions de toutes les valeurs y de la population. L'exposé qui précède vise à démontrer qu'une théorie donnée peut avoir des applications dans divers secteurs pratiques même si elle n'est pas nécessaire dans une situation donnée, comme les enquêtes à grande échelle avec fractions d'échantillonnage du premier degré négligeables. Il montre également que les plans d'échantillonnage avec probabilités inégales jouent un rôle essentiel dans l'échantillonnage d'enquête, malgré l'affirmation de Särndal (1996) selon laquelle des plans simples, comme l'EAS stratifié et l'échantillonnage stratifié de Bernoulli, ainsi que les estimateurs GREG, devraient remplacer les stratégies fondées sur l'échantillonnage avec probabilités inégales sans remise.

Les méthodes-types d'analyse des données sont généralement fondées sur l'hypothèse de l'échantillonnage aléatoire simple, quoique certains progiciels tiennent compte des poids d'échantillonnage et fournissent des estimations ponctuelles correctes. Toutefois, l'application de méthodes-types aux données d'enquête, abstraction faite de l'effet du plan de sondage dû à la mise en grappes et aux probabilités de sélection inégales, risque de produire des inférences erronées, même pour de grands échantillons. En particulier, les erreurs-types des estimations de paramètres et des intervalles de confiance associés peuvent être lourdement sous-estimées, les taux d'erreur de type I des tests d'hypothèses peuvent être beaucoup plus élevés que les niveaux nominaux et les diagnostics de modèles-types, comme l'analyse de résidus pour détecter les écarts par rapport au modèle, sont aussi influencés. Kish et Frankel (1974) et d'autres auteurs se sont penchés sur certains de ces problèmes et ont souligné la nécessité de nouvelles méthodes qui tiennent suffisamment compte de la complexité des données provenant d'enquêtes à grande échelle. Fuller (1975) a mis au point des méthodes asymptotiquement valides d'analyse par régression linéaire, fondées sur des estimateurs de variance par linéarisation de Taylor. Au cours des vingt dernières années, on a fait des progrès rapides en mettant au point des méthodes appropriées. Les méthodes de rééchantillonnage jouent un rôle capital dans la mise au point de méthodes qui tiennent compte du plan d'enquête dans l'analyse des données. On a simplement besoin d'un fichier de données contenant les données observées, des poids d'échantillonnage finaux et des poids finaux correspondant à chaque pseudo-répétition produite par la méthode de rééchantillonnage. On peut alors utiliser des progiciels qui tiennent compte des poids d'échantillonnage dans l'estimation

5. Échantillonnage avec probabilités inégales sans remise

Nous avons mentionné dans la section 2 que l'échantillonnage PPT d'UPÉ à l'intérieur de strates dans les enquêtes à grande échelle était motivé par des considérations pratiques, soit la volonté de répartir des charges de travail à peu près égales. L'échantillonnage PPT permet également de réduire considérablement la variance en ne traitant la variabilité découlant de la taille inégale des UPÉ sans vraiment s'attarder par taille. Les UPÉ sont habituellement échantillonnées sans remise, de manière que la probabilité d'inclusion des UPÉ, π_i , soit proportionnelle à la mesure de la taille des UPÉ x_i . Par exemple, l'échantillonnage PPT systématique, avec ou sans randomisation initiale des étiquettes UPÉ, est un plan avec probabilité d'inclusion proportionnelle à la taille (PPT) (appelé aussi plan πPT) utilisé dans un grand nombre d'enquêtes complexes, dont l'EPA du Canada. L'estimateur d'un total associé à un plan PPT est l'estimateur NHT.

L'élaboration de stratégies appropriées (PIPT, NHT) soulève des problèmes sur le plan théorique, dont l'évaluation de probabilités d'inclusion conjointes exactes, π_{ij} , ou des approximations exactes de π_{ij} nécessitant uniquement les π_i individuels, qui sont nécessaires pour obtenir un estimateur de variance sans biais ou presque sans biais. J'ai étudié ce dernier problème dans la thèse de doctorat que j'ai présentée en 1961 à la Iowa State University. D'éminents statisticiens-mathématiciens ont publié depuis plusieurs solutions nécessitant des outils théoriques perfectionnés. Toutefois, ces travaux théoriques sont souvent qualifiés de « théorie sans application » puisque, dans la pratique, il est courant de traiter les UPÉ comme si elles étaient échantillonnées avec remise, d'où une grande simplification. L'estimateur de variance est obtenu simplement à partir des totaux estimatifs d'UPÉ; cette hypothèse est d'ailleurs à la base des méthodes de rééchantillonnage (section 6). Cet estimateur de variance peut entraîner une surestimation substantielle, sauf si la fraction d'échantillonnage des UPÉ globales est faible, ce qui peut être vrai dans bon nombre d'enquêtes à grande échelle. Dans les paragraphes qui suivent, je tenterai de démontrer que les travaux théoriques portant sur certaines stratégies (PIPT, NHT) et sur des plans de sondage sans PPT ont une grande applicabilité dans la pratique.

J'aborderai d'abord certaines stratégies (PIPT, NHT). En Suède et dans d'autres pays européens, on utilise souvent l'échantillonnage stratifié à un seul degré en raison de la disponibilité de listes et les plans PPT sont des options attrayantes, mais les fractions d'échantillonnage sont souvent grandes. Par exemple, Rosén (1991) mentionne que le baromètre de la population active du Bureau de la statistique

de Suède échantillonne une centaine de populations différentes en utilisant l'échantillonnage PPT systématique et que les taux d'échantillonnage peuvent dépasser 50 %. Aires et Rosén (2005) ont étudié l'échantillonnage πPT de Pareto pour les enquêtes suédoises. Cette méthode possède des propriétés attrayantes, dont la taille fixe de l'échantillon, l'échantillonnage simple, une bonne précision d'estimation, et une estimation convergente de la variance sans égard aux taux d'échantillonnage. En outre, elle permet de coordonner les échantillons au moyen de nombres aléatoires permanents (NAP), comme dans l'échantillonnage de Poisson, mais cette dernière méthode produit des échantillons de taille variable. En raison de ces mérites, on a mis en œuvre l'échantillonnage πPT de Pareto dans un certain nombre d'enquêtes du Bureau de la statistique de Suède, notamment dans les enquêtes sur l'indice des prix. Ohlsson (1995) a décrit les techniques des NAP qui sont couramment utilisées dans la pratique.

La méthode de Rao-Sampford (voir Brewer et Hanif 1983, page 28) produit des plans PPT exacts et des estimateurs de variance non négatifs sans biais pour des échantillons de taille fixe arbitraire. Elle a été mise en œuvre dans la nouvelle version du SAS. Stehman et Overton (1994) notent que la structure de la probabilité variable se manifeste naturellement dans les enquêtes environnementales au lieu d'être sélectionnée uniquement pour l'efficacité accrue et que les π_i sont connus uniquement pour les unités i de l'échantillon s . En traitant le plan de sondage selon la méthode d'échantillonnage systématique aléatoire avec PPT, Stehman et Overton ont obtenu des approximations des π_{ij} qui dépendent uniquement des π_i , $i \in s$, contrairement aux approximations initiales de Hartley et Rao (1962) qui nécessitent la somme des carrés de tous les π_i de la population. Dans les applications de Stehman et Overton, les taux d'échantillonnage sont assez substantiels pour justifier l'évaluation des probabilités d'inclusion conjointes. Je vais maintenant aborder les plans sans PPT utilisant des estimateurs différents de l'estimateur NHT qui assure une variance nulle lorsque y est exactement proportionnel à x . La méthode des groupes aléatoires de Rao, Hartley et Cochran (1962) permet de calculer un estimateur de variance non négatif simple pour n'importe quelle taille fixe de l'échantillon; pourtant, elle se compare favorablement aux stratégies (PIPT, NHT) sur le plan de l'efficacité et elle est toujours plus efficace que la stratégie PPT avec remise. Schabenberger et Grégoire (1994) ont constaté que les stratégies (PIPT, NHT) n'avaient pas trouvé beaucoup d'applications en foresterie à cause de la difficulté de mise en œuvre et ont recommandé la stratégie de Rao-Hartley-Cochran en raison de sa remarquable simplicité et de son efficacité. Il est intéressant de constater que cette stratégie a été utilisée dans l'EPA du Canada parce qu'elle permettait

stables que les poids GREG. Rao et Singh (1997) ont proposé une méthode itérative de « rétrécissement ridge » qui assure la convergence pour un nombre spécifié d'itérations en utilisant une spécification de tolérance intégrée pour assouplir certaines contraintes d'étalement tout en satisfaisant les contraintes relatives à l'étendue. Chen, Sitter et Wu (2002) ont proposé une méthode semblable.

On a utilisé les poids de calage GREG dans l'Enquête sur la population active du Canada qui, tout récemment, a fait appel à des estimateurs composites qui utilisent l'information des mois antérieurs sur l'échantillon, comme nous l'avons mentionné dans la section 2 (Fuller et Rao 2001; Garbino, Kennedy et Singh 2001; Singh, Kennedy et Wu 2001). On a également utilisé des estimateurs par calage de type GREG pour intégrer deux ou plusieurs enquêtes indépendantes portant sur la même population. Ces estimateurs assurent la cohérence entre les enquêtes, en ce sens que les estimateurs de variables communes aux deux enquêtes sont identiques, ainsi que l'étalement en fonction de totaux de population connus (Renssen et Nieuwenbroek 1997; Singh et Wu 1996; Merkouris 2004). Pour le Recensement du Canada de 2001, Bankier (2003) a étudié des poids de calage correspondant à l'estimateur par régression linéaire « optimal » (section 3.3) sous échantillonnage aléatoire stratifié. Il a montré que la méthode de calage « optimale » donnait de meilleurs résultats que l'estimateur par calage GREG, utilisé lors du recensement précédent, dans la mesure où elle permettrait de conserver plus de contraintes d'étalement tout en permettant aux poids de calage d'être au moins un. On peut obtenir le poids de calage « optimal » à l'aide du logiciel SGF en précisant dans les contraintes d'étalement la taille connue des strates et en définissant comme il se doit la constante de réglage q' . Il est à noter que l'estimateur par calage « optimal » possède également des propriétés conditionnelles souhaitables par rapport au plan (section 3.4). Pour la pondération des données du Recensement du Canada de 2001, la méthode de la régression linéaire « optimale » a remplacé celle de l'estimateur GREG par projection (utilisée lors du Recensement de 1996).

Dennett et Rao (2004) ont calculé des estimateurs de variance par linéarisation de Taylor pour une classe générale d'estimateurs par calage avec poids $w_i = d_i / \lambda_i$, où l'on détermine le multiplicateur de Lagrange λ_i en résolvant les contraintes de calage. Le choix $F(a) = 1 + a$ donne des poids GREG et $F(a) = e^a$ permet de calculer des poids obtenus par la méthode itérative du quotient. Dans le cas particulier des poids GREG, l'estimateur de variance se réduit à $v(ge)$ donné dans la section 3.3.

Le lecteur trouvera dans l'article de Fuller (2002), récipiendaire du prix Waksberg, un aperçu et une évaluation plus éloquents de l'estimation par régression dans l'échantillonnage d'enquête, y compris l'estimation par calage.

Il convient de noter que la méthode de rajustement des valeurs de la cellule en fonction des valeurs marginales données dans un tableau à double entrée a d'abord été proposée dans l'article marquant de Deming et Stephan (1940).

Des approches unifiées du calage, fondées sur la minimisation d'une mesure appropriée de la distance entre les poids de calage et les poids de sondage sous réserve des contraintes d'étalement, ont attiré l'attention des utilisateurs en raison de leur capacité de recevoir un nombre arbitraire de contraintes d'étalement spécifiées par l'utilisateur, par exemple, le calage en fonction des valeurs marginales de plusieurs variables de post-stratification. Des logiciels de calage sont également disponibles, dont le SGF (Statistique Canada), LIN WEIGHT (Bureau national de la statistique des Pays-Bas), CALMAR (INSEE, France) et CLAN97 (Bureau de la statistique de Suède).

Une distance de chi carré, $\sum_{i \in q'} (d_i - w_i)^2 / d_i$, permet de calculer l'estimateur GREG (5), où le vecteur $x - w_i(s)$ est dénoté w_i par souci de simplicité (Huang et Fuller 1978; Deville et Särndal 1992). Toutefois, les poids de calage ainsi obtenus ne satisfont pas nécessairement les restrictions relatives à l'étendue souhaitable; par exemple, certains poids peuvent être négatifs ou trop grands, surtout lorsque le nombre de contraintes est élevé et que la variabilité des poids de sondage est élevée. Huang et Fuller (1978) ont proposé une mesure de distance de chi carré modifiée à l'échelle et obtenu les poids de calage au moyen d'une solution itérative qui satisfait les contraintes d'étalement à chaque itération. Toutefois, il n'existe peut-être pas de solution qui satisfait à la fois les contraintes d'étalement et des contraintes relatives à l'ensemble de solutions réalisables satisfaisant les deux types de contraintes peut être vide. D'autres méthodes proposées consistent à modifier la fonction de distance (Deville et Särndal 1992) ou à abandonner certaines contraintes d'étalement (Bankier, Rathwell et Majkowski 1992). Par exemple, une distance d'information de forme $\sum_{i \in q'} \{w_i \log(w_i / d_i) - w_i + d_i\}$ donne des estimateurs de la méthode itérative du quotient avec poids non négatifs w_i , mais certains poids peuvent être beaucoup trop grands. On a également proposé des « poids ridge » obtenus en minimisant une distance de chi carré pénalisée (Chambers 1996), mais rien ne garantit qu'ils satisfassent les contraintes d'étalement ou les contraintes relatives à l'étendue, quoique les poids soient plus

de régression qui fait intervenir la covariance estimative de X_{NHT} et X_{NHT}^* et la variance estimative de X_{NHT}^* . Cet estimateur optimal permet d'établir des inférences conditionnellement valides fondées sur le plan de sondage et est sans biais par rapport au modèle sous le modèle de travail (4). Il s'agit également d'un estimateur par calage dépendant du total X et il peut être exprimé comme suit :

$$\sum_{i \in s} w_i'(s) y_i \text{ avec poids } w_i'(s) = d_i' \hat{g}_i'(s) \text{ et le facteur de calage } \hat{g}_i'(s) \text{ dépendant uniquement du total } X \text{ et les valeurs } x \text{ de l'échantillon. Il fonctionne bien pour l'échantillonage aléatoire stratifié (couramment utilisé dans les enquêtes-établissemements). Toutefois, } B^{opt} \text{ peut devenir instable dans le cas de l'échantillonnage stratifié à plusieurs degrés, sauf si la différence entre le nombre de grappes d'échantillon et le nombre de strates est passablement élevée. L'estimateur GREG n'exige pas cette dernière condition, mais il peut donner de mauvais résultats en ce qui concerne le ratio de biais conditionnel et les taux de couverture conditionnels, comme l'a montré Rao (1996). L'estimateur NHT sans biais peut être conditionnellement très mauvais, sauf si le plan assure que la mesure du déséquilibre définit plus haut est faible. Par exemple, dans le plan de sondage fondé sur la stratification x efficiente et proposé par Hansen et coll. (1983), le déséquilibre est faible et l'estimateur NHT a donné conditionnellement de bons résultats.$$

Tillé (1998) a proposé un estimateur NHT du total Y fondé sur des probabilités d'inclusion conditionnelles approximatives en présence de X_{NHT}^* . Sa méthode permet également d'établir des inférences conditionnellement valides, mais l'estimateur n'est pas calé en fonction de X , contrairement à l'estimateur par régression linéaire « optimal ». Park et Fuller (2005) ont proposé une version calée de l'estimateur GREG fondée sur l'estimateur de Tillé qui donne des poids non négatifs plus souvent que l'estimateur GREG.

Je crois que les praticiens devraient accorder une plus grande attention aux aspects conditionnels de l'inférence fondée sur le plan de sondage et envisager sérieusement les nouvelles méthodes qui ont été proposées.

Kalton (2002) a donné des arguments convaincants pour favoriser des approches fondées sur le plan de sondage (et peut-être conditionnelles ou assistées par un modèle) de l'inférence en fonction des paramètres descriptifs d'une population finie. Smith (1994) a nommé « inférence procédurale » l'inférence fondée sur le plan de sondage et a souligné qu'il s'agissait de l'approche à adopter pour les enquêtes du domaine public. Le lecteur trouvera dans Smith (1976) et Rao et Bellhouse (1990) des études des questions d'inférence dans la théorie des enquêtes par sondage.

On obtient les poids de calage $w_i'(s)$ qui assurent la cohérence avec les totaux auxiliaires X spécifiés par l'utilisateur en rajustant les poids de sondage $d_i' = \pi_i'$ pour satisfaire les contraintes d'échantillonnage $\sum_{i \in s} w_i'(s) x_i = X$. Les estimateurs qui utilisent des poids de calage sont appelés estimateurs par calage et utilisent un seul ensemble de poids $\{w_i'(s)\}$ pour toutes les variables d'intérêt. Nous avons mentionné dans la section 3.4 que l'estimateur GREG assisté par un modèle était un estimateur par calage, mais un estimateur par calage n'est pas nécessairement assisté par un modèle, en ce sens qu'il risque d'être biaisé par rapport au modèle sous un modèle de travail (4), sauf si les variables x du modèle coïncident exactement avec les variables correspondantes aux totaux spécifiés par l'utilisateur. Par exemple, supposons que le modèle de travail suggéré par les données soit un modèle quadratique dans une variable scalaire x alors que le total spécifié par l'utilisateur est uniquement son total X . L'estimateur par calage ainsi obtenu peut donner de mauvais résultats même dans des échantillons assez grands, comme nous l'avons mentionné dans la section 3.3, contrairement à l'estimateur GREG assisté par un modèle fondé sur le modèle quadratique de travail qui nécessite le total de population des variables quadratiques x_i^2 en plus de X .

Dans la pratique, on utilise abondamment la post-stratification pour assurer la cohérence avec les valeurs connues de la cellule correspondant à une variable de post-stratification, par exemple des valeurs dans différents groupes d'âge vérifiées d'après des sources externes comme des projections démographiques. L'estimateur post-stratifié ainsi obtenu est un estimateur par calage. On a également utilisé dans la pratique des estimateurs par calage qui assurent la cohérence avec les valeurs marginales connues de deux ou plusieurs variables de post-stratification, notamment les estimateurs de la méthode itérative du quotient, qu'on obtient par échantillonnage répété des valeurs marginales jusqu'à ce que la convergence soit approximativement réalisée, habituellement en quatre itérations ou moins. Les poids obtenus par la méthode itérative du quotient $w_i'(s)$ sont toujours post-stratifiés. Dans le cadre du Recensement du Canada, Statistique Canada a déjà utilisé les estimateurs de la méthode itérative du quotient pour assurer la cohérence des estimateurs de données-échantillon (2B) avec les valeurs connues des données intégrales (2A). Toujours dans le contexte du Recensement du Canada, Brackstone et Rao (1979) ont étudié l'efficacité des estimateurs de la méthode itérative du quotient et ont aussi calculé des estimateurs de variance par linéarisation de Taylor lorsque le nombre d'itérations était de quatre ou moins. On a également employé les estimateurs de la méthode itérative du quotient dans la Current

$$(8) \quad y_{(f)}^i = x_{(f)}^i + \varepsilon_{(f)}^i, \quad i = 1, \dots, N.$$

Toutefois, les coefficients de régression pondérés ainsi obtenus pourraient devenir instables à cause du risque de multicollinéarité dans l'ensemble de variables auxiliaires. Par conséquent, l'estimateur GREG de $Y^{(1)}$ sous le modèle (8) est moins efficace que l'estimateur GREG sous le modèle (7). En outre, certains poids finaux ainsi obtenus, par exemple $w_i(s)$, risquent de ne pas satisfaire les restrictions relatives à l'étendue en prenant des valeurs inférieures à 1 (dont des valeurs négatives) ou de très grandes valeurs positives. Il est possible de résoudre ce problème en utilisant un estimateur par régression ridge généralisée de $Y^{(1)}$ qui est assisté par un modèle sous le modèle élargi (Chambers 1996; Rao et Singh 1997).

un modèle cherche à utiliser des estimateurs de variance (échantillons) en ce qui concerne la variance conditionnelle par rapport au modèle (du moins pour les grands $v(\lambda)$) dans une notation d'opérateur, un estimateur de variance par linéarisation de Taylor simple satisfaisant la propriété susmentionnée est donné par $v(\lambda e_i)$ où l'on obtient $v(\lambda e_i)$ en remplaçant y_i par $g_i(s_i e_i)$ dans la formule de $v(\lambda)$; voir Hidiogrou, Fuller et Hickman (1976) et Sämndal, Svensson et Wretman (1989).

Dans l'exposé qui précède, nous avons supposé un modèle de régression linéaire de travail pour toutes les variables $y^{(j)}$. Dans la pratique, cependant, un modèle de régression linéaire n'est pas nécessairement bien adapté à certaines variables d'intérêt y , par exemple, une variable binaire. Dans ce dernier cas, la régression logistique offre un modèle de travail approprié. Un modèle de travail général qui couvre la régression logistique prend la forme $E^m(y_i) = h(x_i'\beta) = \mu_i$, où $h(\cdot)$ pourrait être non linéaire: le modèle (5) est un cas particulier avec $h(a) = a$. Un estimateur basé sur un modèle de travail sous la forme de travail général est l'estimateur par la différence $X_{NHT}^{NHT} + \sum_i u_i \mu_i - \sum_i \pi_i^1 \mu_i$, où $\mu_i = h(x_i'\beta)$ et β est un estimateur du paramètre de modélisation β . Il se réduit à l'estimateur GREG (5) si $h(a) = a$. Cet estimateur par la différence est presque optimal si la probabilité d'inclusion π_i est proportionnelle à σ_i , où σ_i^2 dénote la variance par rapport au modèle, $V^m(y_i)$.

Les estimateurs GRECQ sont très appréciés par les utilisateurs parce que bon nombre d'estimateurs couramment utilisés peuvent être obtenus sous forme de cas particuliers de (5) par des spécifications appropriées de x_1 et q . Statistique Canada a mis au point un Système généralisé d'estimation (SGE) fondé sur l'estimateur GRECQ.

Kott (2005) a proposé un autre paradigme de l'inférence, appelé approche fondée sur un modèle et assistée par randomisation, qui est axée sur l'inférence fondée sur un modèle et assistée par randomisation (ou échantillonnage répété). La définition de la variance prévue est inversée pour devenir la variance prévue par rapport au modèle à randomisation d'un estimateur, mais elle est identique à la variance prévue habituelle lorsque le modèle de travail est valable pour l'échantillon, comme on le suppose dans l'article. Par conséquent, les choix de l'estimateur et de l'estimateur de variance sont souvent semblables à ceux qui sont faits sous l'approche assistée par un modèle. Toutefois, Kott soutient que la motivation est plus claire et que « l'approche populationnelle pour l'estimation de la variance mène, au besoin, à un traitement logiquement cohérent de rajustements d'une population finie et d'un petit échantillon ».

3.4 Approche conditionnelle fondée sur le plan de sondage

On a également proposé une approche conditionnelle fondée sur le plan de sondage. Cette approche cherche à combiner les caractéristiques conditionnelles de l'approche dépendante d'un modèle avec les caractéristiques indépendantes de l'approche fondée sur le plan de sondage. Elle permet de restreindre l'ensemble d'échantillons de référence à un sous-ensemble « pertinent » de tous les échantillons possibles spécifiés par le plan de sondage. On obtient des inférences conditionnellement valides en ce sens que le ratio de biais conditionnel (soit le ratio du biais conditionnel à l'erreur-type conditionnelle) devient nul à mesure que la taille de l'échantillon augmente. Environ $100(1 - \alpha) \%$ des intervalles de confiance réalisés dans l'échantillonnage répétée à partir de l'ensemble conditionnel contiennent le total inconnu Y .

Holt et Smith (1979) fournissent des arguments convaincants

plans en faveur de l'inférence conditionnelle fondée sur le simple d'un échantillon aléatoire simple, auquel cas l'est naturel de faire des inférences conditionnelles à la taille des strates de l'échantillon réalisé. Rao (1992, 1994) et Casady et Valliant (1993) ont étudié l'inférence conditionnelle lorsque seul le total auxiliaire X est connu d'après des sources externes. Dans ce dernier cas, la subordination à l'estimateur NHT X^{NHT} peut s'avérer raisonnable parce qu'il s'agit « à peu près » d'une statistique auxiliaire lorsque X est connu et que la différence $X^{NHT} - X$ fournit une mesure du déséquilibre de l'échantillon réalisé. La subordination à X^{NHT} permet de calculer l'estimateur par régression linéaire « optimal », de même forme que l'estimateur GREG (5), dans lequel B donné par (6) est remplacé par la valeur optimale estimative B_0^* du coefficient

représentation montre le rôle du modèle de travail dans l'approche assistée par un modèle. L'estimateur GREG (5) est cohérent avec le plan de sondage et sans biais par rapport au modèle sous le travail, à condition que la probabilité d'inclusion, π_i , soit proportionnelle à l'écart-type par rapport au modèle σ_i . Toutefois, dans les enquêtes à plusieurs variables d'intérêt, la variance par rapport au modèle peut varier selon les variables. Comme on doit utiliser un plan de sondage général, tel que le plan avec probabilités d'inclusion proportionnelles à la taille, le résultat d'optimalité n'est plus valable, même si le même vecteur x_i est utilisé pour toutes les variables y_i du modèle de travail.

L'estimateur GREG devient simplement l'estimateur « par projection » $X'B = \sum w_i(s)y_i$ avec $g_i(s) = X'T^{-1}x_i/q_i$ si la variance par rapport au modèle σ_i^2 est proportionnelle à $\lambda'x_i$ pour certains λ . On obtient l'estimateur par quotients sous forme de cas particulier de l'estimateur par projection, étant donné $g_i = x_i$, d'où $g_i(s) = X/X^{HT}$. Il est à noter que l'estimateur GREG (5) exige sairement les valeurs de population individuelles x_i . Cette caractéristique est très utile, car les totaux de population auxiliaire sont souvent attestés à l'aide de sources externes comme les projections démographiques des dénombremenets selon l'âge et le sexe. De plus, il assure la cohérence avec les totaux connus X en ce sens que $\sum w_i(s)x_i = X$. En raison de cette propriété, l'estimateur GREG est également un estimateur par calage.

Supposons qu'il y ait p variables d'intérêt, par exemple $y_{(1)}, \dots, y_{(p)}$, et qu'on veuille utiliser l'approche assistée par un modèle pour estimer les totaux de population correspondants $Y_{(1)}, \dots, Y_{(p)}$. Supposons également que le modèle de travail de $y_{(f)}$ prenne la forme (4) mais nécessairement un vecteur x peut-être différent $x_{(f)}$ avec total connu $X_{(f)}$ pour chaque $f = 1, \dots, p$:

$$y_{(f)}^i = (x_{(f)}^i)\beta_{(f)} + \varepsilon_{(f)}^i, \quad i = 1, \dots, N. \quad (7)$$

Dans ce cas, les poids g dépendent de f et, à leur tour, les poids finaux $w_i(s)$ dépendent aussi de f . Dans la pratique, il est souvent souhaitable d'utiliser un seul ensemble de poids finaux pour toutes les variables p afin d'assurer la cohérence interne des chiffres lorsqu'ils sont agrégés à partir de variables différentes. On ne peut réaliser cette propriété qu'en élargissant le vecteur x dans le modèle (7) pour recevoir toutes les variables $y_{(f)}$, par exemple \tilde{x} avec total connu \tilde{X} , puis en utilisant le modèle de travail

donnent de bons résultats. Par contre, pour des populations très asymétriques, les intervalles normaux fondés sur X^{NHT} et son erreur-type peuvent donner de mauvais résultats sous échantillonnage répété, même pour des échantillons assez grands, car le pivot dépend de l'asymétrie des y_i . La structure de population joue donc un rôle dans les inférences fondées sur le plan de sondage, contrairement à ce qu'affirment Neyman (1934), Hansen et coll. (1983) et d'autres auteurs. Rao, Jocelyn et Hidiroglou (2003) ont considéré l'estimateur par régression linéaire simple sous échantillonnage aléatoire simple à deux phases avec seulement x observé dans la première phase. Ils ont démontré que le rendement de couverture des intervalles normaux associés pouvait être faible même pour des échantillons de deuxième phase passablement grands si le vrai modèle sous-jacent qui produit la population s'écarterait considérablement du modèle de régression linéaire (par exemple, une régression quadratique de y sur x) et si l'asymétrie de x est grande.

Dans ce cas, les valeurs x de la première phase sont observées et une approche assistée par un modèle approprié utiliserait un estimateur par régression linéaire multiple avec x et $z = x^2$ comme variables auxiliaires. Il est à noter que pour l'échantillonnage à une seule phase, on ne peut mettre en œuvre un tel estimateur assisté par un modèle si l'on connaît uniquement les X , puisque l'estimateur dépend du total de population de z .

Särndal, Swenson et Wretman (1992) proposent une description complète de l'approche assistée par un modèle pour estimer le total Y d'une variable y sous le modèle de régression linéaire de travail

$$y_i = x_i'\beta + \varepsilon_i; \quad i = 1, \dots, N \quad (4)$$

avec moyenne nulle, erreurs non corrélées ε_i et variance par rapport au modèle $V_m(\varepsilon_i) = \sigma^2 q_i = \sigma^2_i$ où les q_i sont des constantes connues et les vecteurs x ont des totaux connus X (les valeurs de population x_1, \dots, x_N ne sont pas nécessairement connues). Dans ces conditions, l'approche assistée par un modèle produit l'estimateur de régression généralisée (generalized regression ou GREG)

$$\hat{Y}_{gr} = \hat{Y}^{NHT} + B'(X - X^{NHT}) = \sum_{i \in s} w_i(s)y_i \quad (5)$$

où

$$B = T^{-1} \left(\sum^s \pi_i^{-1} x_i y_i / q_i \right) \quad (6)$$

avec $T = \sum^s \pi_i^{-1} x_i x_i' / q_i$ est un coefficient de régression pondéré et $w_i(s) = g_i(s) \pi_i^{-1}$ avec $g_i(s) = 1 + (X - X^{NHT})' T^{-1} x_i / q_i$, appelé « poids g ». Il est à noter que l'estimateur GREG (5) peut également s'écrire $\sum^s \tilde{w}_i y_i + \hat{Y}^{NHT}$, où \hat{Y}^{NHT} est le prédicteur de Y_i sous le modèle de travail et \hat{E}^{NHT} est l'estimateur NHT de l'erreur de prévision totale $E = \sum^s e_i$ avec $e_i = y_i - \hat{y}_i$. Cette

3.2 Approche dépendante d'un modèle

En matière d'inférence, l'approche dépendante d'un modèle suppose que la structure de population obéit à un modèle de superpopulation spécifié. La distribution induite par le modèle hypothétique produit des inférences qui renvoient à l'échantillon donné d'unités s qui a été tiré. Ces inférences conditionnelles peuvent s'avérer plus pertinentes et plus attrayantes que les inférences établies par échantillonnage répété. Par contre, lorsque le modèle n'est pas spécifié correctement, les stratégies dépendantes d'un modèle peuvent donner de mauvais résultats dans le cas de grands échantillons; même de faibles écarts par rapport au modèle hypothétique, difficiles à déceler au moyen de méthodes de vérification de modèle, peuvent causer de graves problèmes. Par exemple, prenons le modèle par quotient souvent utilisé lorsqu'une variable auxiliaire x au total connu X est aussi mesurée dans l'échantillon :

$$(2) \quad y_i = \beta x_i + \varepsilon_i; i = 1, \dots, N$$

où les ε_i sont des variables aléatoires indépendantes avec moyenne nulle et variance proportionnelle à x_i . En supposant que le modèle soit valable pour l'échantillon, c'est-à-dire sans biais d'échantillonnage, le meilleur prédicteur sans biais par rapport au modèle linéaire du total X est donné par l'estimateur par quotient $(\bar{y}/\bar{x}) X$ sans égard au plan de sondage. Cet estimateur n'est pas convergent selon le plan de sondage, sauf si le plan est autopondéré, par exemple sous échantillonnage aléatoire stratifié avec répartition proportionnelle. Par conséquent, sous des plans non autopondérés, il peut donner de très mauvais résultats dans le cas de grands échantillons, même si les écarts par rapport au modèle sont faibles. Hansen et coll. (1983) ont démontré les mauvais résultats obtenus dans des conditions d'échantillonnage répété, en utilisant un plan d'échantillonnage aléatoire stratifié avec une répartition de l'échantillon presque optimale (couramment utilisé en présence de populations très asymétriques). Rao (1996) a utilisé le même plan pour démontrer les mauvais résultats obtenus dans le contexte d'un cadre conditionnel pertinent à l'approche dépendante d'un modèle (Royall et Cumberland 1981). Néanmoins, les approches dépendantes d'un modèle peuvent jouer un rôle capital dans l'estimation sur petits domaines où la taille de l'échantillon dans un petit domaine peut être infime, voire nulle (voir la section 7).

Brewer (1963) a été le premier à proposer l'approche dépendante d'un modèle dans le contexte du modèle par quotient (2). Royall (1970) et ses collaborateurs ont mené une étude systématique de cette approche. Valliant, Dorfman et Royall (2000) donnent une description complète de la théorie, dont l'estimation de la variance (conditionnelle) par rapport au modèle de l'estimateur qui varie avec s ; par exemple, sous le modèle par quotient (2), la variance par

3.3 Approche assistée par un modèle

L'approche assistée par un modèle cherche à combiner les caractéristiques positives de la méthode fondée sur le plan de sondage et de la méthode dépendante d'un modèle. Elle considère uniquement les estimateurs convergents selon le plan de sondage du total X qui sont aussi sans biais par rapport au modèle sous le modèle « de travail » hypothétique. Par exemple, sous le modèle par quotient (2), un estimateur assisté par un modèle de X pour un plan d'échantillonnage probabiliste spécifié est donné par l'estimateur par quotient $\hat{Y}_p = (\hat{Y}^{NHT}/\hat{X}^{NHT}) X$ qui est convergent selon le plan de sondage sans égard au modèle hypothétique. Hansen et coll. (1983) ont utilisé cet estimateur dans leur plan d'échantillonnage stratifié pour démontrer que ses résultats étaient supérieurs à ceux de l'estimateur dépendant d'un modèle $(\bar{y}/\bar{x}) X$. Pour l'estimation de la variance, l'approche assistée par un modèle utilise des estimateurs convergents pour la variance de l'estimateur par rapport au plan tout en étant exactement ou asymptotiquement sans biais par rapport au modèle pour la variance par rapport au modèle. Toutefois, les inférences sont fondées sur le plan de sondage, car le modèle est utilisé uniquement comme modèle « de travail ».

Pour l'estimateur par quotient \hat{Y}_p , l'estimateur de variance est donné par

$$(3) \quad \text{Var}(\hat{Y}_p) = (X/\hat{X}^{NHT})^2 v(e),$$

où, dans la notation d'opérateur, $v(e)$ est obtenu à partir de $v(y)$ en remplaçant y_i par les résidus $e_i = y_i - (\hat{Y}^{NHT}/\hat{X}^{NHT}) x_i$. Cet estimateur de variance est asymptotiquement équivalent à un estimateur de linéarisation courant de la variance $v(e)$, mais il reflète le fait que l'information contenue dans l'échantillon varie avec \hat{X}^{NHT} : les valeurs élevées produisent une faible variabilité, et les valeurs faibles, une grande variabilité. Le pivot normal ainsi obtenu produit des inférences dépendantes d'un modèle qui sont valides sous le modèle hypothétique (contrairement à l'utilisation de $v(e)$ dans le pivot) tout en protégeant contre les écarts par rapport au modèle, en ce sens qu'il produit des inférences asymptotiquement valides fondées sur le plan de sondage. Il est à noter que le pivot est asymptotiquement équivalent à $Y(\hat{e})/[v(\hat{e})]^{1/2}$ avec $\hat{e}_i = y_i - (Y/X) x_i$. Si dans les résidus \hat{e}_i est faible même si y_i et x_i sont très asymétriques, et les intervalles de confiance normaux

informative de la fonction de vraisemblance est due à la propriété d'étiquette qui traite les unités de population N essentiellement comme des post-strates N . On peut contourner cette difficulté en employant la méthode bayésienne et en supposant des valeurs antérieures informatives (échangeables) sur le vecteur paramètre (Ericson 1969). Une autre solution (fondée sur le plan de sondage) consiste à faire abstraction de certains aspects des données-échantillons pour rendre l'échantillon non unique et arriver ainsi à une fonction de vraisemblance informative (Hartley et Rao 1968; Royall 1968). Par exemple, sous échantillonnage aléatoire simple, en supprimant les étiquettes i et en considérant les données $\{(t, y_i), i \in s\}$ en l'absence d'information liant i à y_i , on obtient la moyenne d'échantillon comme estimateur du maximum de vraisemblance de la moyenne de population. En supposant des distributions antérieures non informatives, l'estimation bayésienne produit des résultats semblables à ceux obtenus par Ericson (1969) mais, contrairement à l'estimation d'Ericson, elle dépend du plan de sondage. Dans le cas où y_i est un vecteur qui comprend des variables auxiliaires avec totaux connus, Hartley et Rao (1968) ont montré que sous échantillonnage aléatoire simple, l'estimateur du maximum de vraisemblance était à peu près égal à l'estimateur par régression traditionnel du total. Cet article a été le premier à montrer comment intégrer des totaux de population auxiliaire connus à un cadre de vraisemblance. Pour l'échantillonnage aléatoire stratifié, on fait abstraction des étiquettes à l'intérieur des strates, mais pas des étiquettes de strate, à cause de différences connues entre les strates. L'estimateur du maximum de vraisemblance ainsi obtenu est à peu près égal à un pseudo-estimateur par régression linéaire optimal lorsqu'on dispose de variables auxiliaires avec totaux connus. Ce dernier estimateur possède de bonnes propriétés conditionnelles fondées sur le plan de sondage (voir la section 3.4). L'article de Hartley et Rao (1968) portait sur l'estimation d'un total, mais l'approche de la vraisemblance a une portée beaucoup plus vaste en échantillonnage, dont l'estimation de fonctions de distribution et de quantiles et la construction d'intervalles de confiance fondés sur des rapports de vraisemblance (voir la section 8.1). L'approche de la vraisemblance non paramétrique de Hartley-Rao a été découverte indépendamment vingt ans plus tard (Owen 1988) dans l'inférence statistique courante, sous le nom de « vraisemblance empirique », et a attiré passablement d'attention, notamment pour son application à divers problèmes d'échantillonnage. Dans un certain sens, les efforts d'intégration à la statistique courante ont donc partiellement réussi. L'ouvrage d'Owen (2002) présente une description complète de la théorie de la vraisemblance empirique et de ses applications.

On a aussi tenté d'intégrer la théorie des enquêtes par sondage à l'inférence statistique courante au moyen de la fonction de vraisemblance. Godambe (1966) a montré que la fonction de vraisemblance d'après les données-échantillons $\{(t, y_i), i \in s\}$, en considérant comme paramètre le vecteur N des valeurs y inconnues, ne fournissait pas d'information sur les valeurs non observées de l'échantillon ni, par conséquent, sur le total Y . Cette caractéristique non

même sous échantillonnage aléatoire simple. Ce résultat théorique négatif importait à être, dans une grande mesure, négligé pendant une dizaine d'années. Godambe a également établi un résultat positif en liant y à une mesure de taille x au moyen d'un modèle de régression de superpopulation passant par l'origine avec variance d'erreur proportionnelle à x^2 , puis en montrant que l'estimateur NHT sous un plan de sondage à taille fixe où π_i est proportionnel à x_i minimisait la variance prévue de la classe sans biais (1). Ce résultat montre clairement les conditions du plan pour l'utilisation de l'estimateur NHT. Rao (1966) a constaté les limites de l'estimateur NHT dans le contexte d'enquêtes avec échantillonnage PPT et caractéristiques multiples. Ici, l'estimateur NHT s'avère très inefficent lorsqu'une caractéristique y n'est pas liée (ou qu'elle est faiblement liée) à la mesure de taille x (comme le dénombrement de volailles y et la taille de la ferme x dans une enquête sur les fermes). Rao a proposé pour ces cas d'autres estimateurs efficaces qui font abstraction des poids NHT. En faisant abstraction des résultats susmentionnés, des spécialistes de l'échantillonnage ont avancé plus tard certains critères théoriques pour affirmer qu'il fallait utiliser l'estimateur NHT pour tout plan de sondage. En prenant l'exemple amusant des éléphants d'un cirque, Basu (1971) a illustré la futilité de ces critères. Il a construit un « mauvais » plan dans lequel π_i n'était pas lié à y_i pour démontrer que l'estimateur NHT produisait des estimations absurdes, ce qui a incité le célèbre statisticien bayésien Dennis Lindley à conclure que ce contre-exemple détruisait la théorie des enquêtes par sondage fondées sur le plan de sondage (Lindley 1996). Cette conclusion est plutôt malheureuse, car NHT et Godambe ont clairement énoncé les conditions du plan pour une utilisation appropriée de l'estimateur NHT, et Rao (1966) et Hajek (1971) ont proposé d'autres estimateurs pour composer respectivement avec les caractéristiques multiples et les mauvais plans. Il est intéressant de noter que les mêmes critères théoriques ont abouti à un mauvais choix « optimal » (Rao et Singh 1973).

$$(1) \quad \bar{y} = \sum_{i \in s} d_i^{-1}(y) y_i$$

c'est-à-dire un poids de forme $d_i^{-1}(y)$. Il a alors établi que l'estimateur BLUE n'existait pas dans la classe générale

On a envisagé des stratégies (plan de sondage et estimation) qui semblaient raisonnables et l'on a soigneusement étudié des propriétés relatives au moyen de méthodes analytiques ou empiriques, et parfois aussi des erreurs quadratiques moyennes ou des variances prévues sous des modèles de superpopulation plausibles, comme nous le mentionnons dans la section 2. On n'a pas insisté sur une estimation sans biais sous un plan de sondage donné, car elle « entraîne souvent une erreur quadratique moyenne beaucoup plus grande que nécessaire » (Hansen, Hurvitz et Tepping 1983). On a plutôt jugé que la cohérence avec le plan de sondage était nécessaire pour les grands échantillons. Les ouvrages classiques de Cochran (1953), Deming (1950), Hansen, Hurvitz et Madow (1953), Sukhatme (1954) et Yates (1949), fondés sur l'approche susmentionnée, ont grandement influencé la pratique des enquêtes. Pourtant, les statisticiens universitaires accordaient peu d'attention à la théorie de l'échantillonnage traditionnelle, peut-être parce qu'il lui manquait un cadre théorique formel et qu'elle n'était pas intégrée à la théorie statistique courante. Plusieurs départements de statistique nord-américains de prestige n'offraient pas de cours supérieurs en théorie de l'échantillonnage.

Dans les années 1950, on a élaboré des cadres et des approches théoriques formels pour intégrer la théorie de l'échantillonnage à l'inférence statistique courante dans des conditions quelque peu idéalistes axées sur les erreurs d'échantillonnage, en supposant l'absence d'erreurs de mesure ou de réponse ainsi que de non-réponse. Horvitz et Thompson (1952) ont apporté une contribution de base à l'échantillonnage avec probabilités de sélection arbitraires en formulant trois sous-classes d'estimateurs linéaires sans biais d'un total X , dont la classe de Markov étudiée par Neyman. Une autre sous-classe avec poids de sondage d_i liée à une unité d'échantillonnage i et dépendant uniquement de i admettait l'estimateur bien connu avec poids inversement proportionnel à la probabilité d'inclusion π_i comme seul estimateur sans biais. Narain (1951) ayant également découvert cet estimateur, on devrait l'appeler l'estimateur de Narain-Horvitz-Thompson (NHT) au lieu de l'estimateur HT comme on l'appelle couramment. Pour l'échantillonnage aléatoire simple, la moyenne d'échantillon est le meilleur estimateur linéaire sans biais (best linear unbiased estimator ou BLUE) de la moyenne de population dans les trois sous-classes, mais ce n'est pas suffisant pour prétendre que la moyenne d'échantillon est le meilleur de tous les estimateurs linéaires sans biais. Godambe (1955) a proposé une classe générale d'estimateurs linéaires sans biais d'un total X en supposant des données-échantillons $\{(i, y_i), i \in s\}$ et un poids dépendant de l'unité d'échantillonnage i ainsi que des autres unités échantillonnées s ,

Au départ, l'élaboration de la théorie de l'échantillonnage a progressé de manière plus ou moins inductive, quoique Neyman (1934) ait étudié la meilleure estimation linéaire sans biais pour l'échantillonnage aléatoire stratifié.

3.1 Cadre unifié fondé sur le plan de sondage

3. Questions d'inférence

Le concept de l'effet du plan de sondage (EPS), dû à Leslie Kish (voir Kish 1965, section 8.2), constitue un autre jalon de la méthodologie des enquêtes par sondage. L'effet statistique sous le plan de sondage spécifié à la variance qui serait obtenue sous échantillonnage aléatoire simple de même taille. Ce concept est particulièrement utile dans la présentation et la modélisation des erreurs d'échantillonnage, ainsi que dans l'analyse des données d'enquête complexes faisant intervenir la mise en grappes et les probabilités de sélection inégales (voir la section 6).

Le lecteur trouvera dans Kish (1995), Kruskal et Mosteller (1980), Hansen, Dalenius et Tepping (1985) et O'Muircheartaigh et Wong (1981) un examen des apports marquants à la théorie et aux méthodes des enquêtes par sondage.

La variance totale qui tient compte de la variance de réponse corrélée due aux intervieweurs. Les sous-échantillons superposés entraînent une augmentation des frais de déplacement des intervieweurs, mais on peut les réduire en modifiant les affectations des intervieweurs. Hansen, Hurvitz, Marks et Mauldin (1951), Sukhatme et Seth (1952) et Hansen, Hurvitz et Bershad (1961) ont formulé des théories de base sous des modèles d'erreur de mesure additive et décomposée la variance totale en trois éléments : la variance d'échantillonnage, la variance de réponse simple et la variance de réponse corrélée. On a montré que la variance de réponse corrélée due aux intervieweurs était de l'ordre de k^{-1} sans égard à la taille de l'échantillon, k étant le nombre d'intervieweurs. Par conséquent, elle peut dominer la variance totale si k n'est pas un nombre élevé. Lors du recensement de 1950 aux États-Unis, l'étude de la variance due aux intervieweurs a montré que cette composante était effective-ment grande pour les petits domaines. C'est en partie pour cette raison que lors du recensement de 1960, on a adopté l'autodénombrément par la poste pour réduire cette composante de la variance (Waksberg 1998). Il s'agit d'un exemple éloquent de l'influence de la théorie sur la pratique. Fellegi (1964) a proposé de combiner la superposition et la répétition pour estimer la covariance entre l'écart d'échantillonnage et l'écart de réponse. Cette composante est souvent négligée dans la décomposition de la variance totale, mais elle pourrait être appréciable dans la pratique.

régression, qui utilise l'information sur l'échantillon obtenue au cours des mois précédents et qui peut être mise en œuvre avec un logiciel de poids de régression (voir la section 4).

Keyfitz (1951) a proposé une méthode ingénieuse pour obtenir de meilleures mesures de la taille des UPE dans les enquêtes permanentes fondées sur les plus récents dénombrements censitaires. Sa méthode permet de maximiser la probabilité de chevauchement avec l'échantillon antérieur d'une UPE par strate, ce qui réduit les coûts d'opération sur le terrain tout en améliorant l'efficacité grâce aux meilleures mesures de la taille dans l'échantillonnage PPT. L'EPA du Canada et d'autres enquêtes permanentes ont utilisé la méthode de Keyfitz. Raj (1956) a formulé le problème de l'optimisation comme un « problème de transport » dans la programmation linéaire. Kish et Scott (1971) ont étendu la méthode de Keyfitz aux mesures changeantes des strates et de la taille. Ernst (1999) a proposé un excellent tableau de l'évolution, au cours des 50 dernières années, de la coordination d'échantillons (qui consiste à maximiser ou minimiser le chevauchement des échantillons) au moyen d'algorithmes de transport et de méthodes connexes; voir aussi Mach, Reiss et Schiopu-Kratina (2005) en ce qui concerne les applications aux enquêtes-entreprises avec création et suppression d'entreprises.

Dalenius (1957, chapitre 7) a étudié le problème de la stratification optimale d'un nombre donné de strates, L , dans le cadre de la répartition de Neyman. Dalenius et Hodges (1959) ont obtenu une approximation simple de la stratification optimale, appelée méthode de la fonction cumulative de la racine carrée des fréquences (cum \sqrt{f}), qui est abondamment utilisée dans la pratique. Pour les populations très asymétriques dont un petit nombre d'unités comptent pour une forte proportion du total X , comme les populations d'entreprises, une stratification efficace nécessite une strate à tirage complet ($n_1 = N_1$) de grandes unités et des strates à tirage partiel d'unités moyennes et petites. Lavallée et Hidiroglou (1988) et Rivest (2002) ont mis au point des algorithmes pour déterminer les bornes de stratification en utilisant la méthode puissance (Fellegi 1981; Bankier 1988) et la répartition de Neyman pour les strates à tirage partiel. Aujourd'hui, Statistique Canada et d'autres organismes utilisent ces algorithmes pour les enquêtes-entreprises.

Avant 1950, la recherche portant sur l'estimation de taux et de moyennes de population pour la population entière et de grandes sous-populations planifiées, comme des États ou des provinces. Or, les utilisateurs s'intéressent également aux taux et aux moyennes de sous-populations non planifiées (appelées aussi domaines), comme les groupes d'âge-sexe à l'intérieur d'une province, ainsi qu'à des paramètres autres que les totaux et les moyennes,

comme les médianes et d'autres quantiles, par exemple le revenu médian. Hartley (1959) a formulé une théorie simple et unifiée de l'estimation par domaine, applicable à n'importe quel plan de sondage et nécessitant uniquement les formules-types pour l'estimateur du total et son estimateur de variance, dénotés respectivement $X(y)$ et $v(y)$ dans la notation d'opérateur. Il a introduit deux variables synthétiques y_i et a_i qui apparaissent respectivement les valeurs y_i et a_i si l'unité i appartient au domaine j et qui sont nulles dans le cas contraire. Alors, on obtient simplement les estimateurs du total de domaine j $X = X(j, y)$ et de la taille de domaine j $N = X(j, a)$ à l'aide des formules pour $X(y)$ et $v(y)$ en remplaçant respectivement y_i par j, y_i et a_i . De même, on obtient les estimateurs des moyennes de domaine et des différences de domaine ainsi que leurs estimateurs de variance à l'aide des formules de base pour $X(y)$ et $v(y)$. Durbin (1968) a également obtenu des résultats semblables. Aujourd'hui, on pratique couramment l'estimation par domaine en utilisant l'ingénieuse méthode de Hartley.

Pour l'inférence concernant des quantités, Woodruff (1952) a proposé une méthode simple et ingénieuse pour obtenir un intervalle de confiance de niveau $(1 - \alpha)$ sous des plans d'échantillonnage généraux, en utilisant uniquement la fonction de distribution estimative et son erreur-type (voir l'ouvrage de Lohr (1999), pages 311 à 313). Il est à noter qu'on obtient simplement ces dernières à l'aide des formules pour un total en remplaçant y par une variable indicatrice. En mettant sur le même pied l'intervalle de Woodruff et un intervalle selon la théorie normale à l'égard du quantile, on peut aussi obtenir une formule simple pour l'erreur-type du d^* estimateur de quantile, soit la moitié de la longueur de l'intervalle divisé par le point supérieur $\alpha/2$ de la distribution normalisée $N(0, 1)$ qui égale 1,96 si $\alpha = 0,05$ (Rao et Wu 1987; Francisco et Fuller 1991). Il donne de bons résultats même lorsque d^* est petit ou grand et que la taille de l'échantillon est moyenne (Sitter et Wu 2001).

On s'est rendu compte de l'importance des erreurs de mesure dès les années 1940. Dans un article influent, Mahalanobis (1946a) a mis au point la technique des sous-échantillons superposés (appelée échantillonnage répété par Deming 1960). En Inde, on a beaucoup utilisé cette méthode dans les enquêtes par sondage à grande échelle pour évaluer les erreurs d'échantillonnage et les erreurs de mesure. L'échantillon est tiré sous forme de deux ou plusieurs sous-échantillons indépendants selon le même plan de sondage, de sorte que chaque sous-échantillon fournit une estimation valide du total ou de la moyenne. Les sous-échantillons sont attribués à des intervieweurs différents (ou à des équipes différentes), ce qui produit une estimation valide de la

article marquant (Mahalanobis 1944) présente des résultats théoriques probants sur la conception efficace d'enquêtes par sondage et leurs applications pratiques, notamment dans le cas d'enquêtes sur les surfaces cultivées et le rendement des cultures. Maintenant bien connue, la répartition optimale sous échantillonnage aléatoire stratifié où le coût par unité varie d'une strate à l'autre est obtenue sous forme de cas particulier de sa théorie générale. Dès 1937, Mahalanobis a utilisé des plans de sondage à plusieurs degrés pour les enquêtes sur le rendement des cultures avec, comme unités d'échantillonnage aux quatre degrés d'échantillonnage, des villages, des grilles à l'intérieur des villages, des parcelles à l'intérieur des grilles et des coupes de tailles et de formes différentes (Murthy 1964). Il a également utilisé un plan de sondage à deux phases pour estimer le rendement de l'écorce de quinquina. Il a joué un rôle de premier plan dans l'établissement de la National Sample Survey (NSS) de l'Inde, la plus vaste enquête polyvalente permanente : un personnel à temps plein effectue des interviews sur place pour des enquêtes socioéconomiques et des mesures physi-ques pour des enquêtes sur les cultures. Plusieurs éminents statisticiens d'enquête, dont D.B. Lahiri et M.N. Murthy,

ont collaboré à la NSS.

P.V. Sukhatme, qui a étudié avec Neyman, a également fait un apport innovateur à la conception et à l'analyse d'enquêtes agricoles à grande échelle en Inde, en utilisant l'échantillonnage stratifié à plusieurs degrés. À partir de 1942-1943, il a mis au point des plans de sondage efficaces pour mener des enquêtes nationales sur les cultures de blé et de riz et a obtenu un degré élevé de précision pour les estimations nationales ainsi qu'une marge d'erreur raisonnable pour les estimations par district. L'approche de Sukhatme différait de celle de Mahalanobis, qui utilisait des parcelles de très petite taille pour les coupes-témoins et employait des enquêteurs *ad hoc*. Sukhatme (1947) et Sukhatme et Panse (1951) ont démontré que l'utilisation d'une petite parcelle pourrait donner des estimations biaisées à cause de la tendance à placer des plantes de boma-ges à l'intérieur de la parcelle lorsque l'on doute. Ils ont également souligné que le recours à des enquêteurs *ad hoc*, qui se déplacent rapidement d'un endroit à l'autre, obligerait à mesurer uniquement les parcelles de champs échantillonnées qui sont prêts à moissonner à la date de la visite, ce qui est contraire au principe de l'échantillonnage aléatoire. La solution de Sukhatme consistait à utiliser de grandes parcelles pour éviter les biais liés au boma-ges et à compter les coupes-témoins à l'organisme public local chargé du revenu ou de l'agriculture.

De 1940 à 1970, les statisticiens d'enquête du U.S. Census Bureau, sous la direction de Morris Hansen, William Hurwitz, William Madow et Joseph Waksberg, ont fait des apports fondamentaux à la théorie et à la pratique des

enquêtes par sondage, et bon nombre de leurs méthodes sont encore largement utilisées dans la pratique. Hansen et Hurwitz (1943) ont formulé la théorie de base de l'échantillonnage stratifié à deux degrés, une seule unité primaire d'échantillonnage (UPÉ) à l'intérieur de chaque strate étant tirée avec probabilité proportionnelle à la taille (échantillonnage PPT) puis sous-échantillonnée à un rythme qui assure l'autopondération (probabilités de sélection globales égales) à l'intérieur des strates. Cette approche permet de confier aux intervieweurs des charges de travail à peu près égales, ce qui est souhaitable dans le contexte des enquêtes sur le terrain. Elle permet aussi de réduire considérablement la variance en neutralisant la variabilité due à la taille inégale des UPÉ sans vraiment stratifier selon la taille, ce qui permet la stratification selon d'autres variables pour réduire la variance. Par contre, les charges de travail peuvent varier considérablement si les UPÉ sont sélectionnées par échantillonnage aléatoire simple, puis sous-échantillonnées au même rythme à l'intérieur de chaque strate. Aujourd'hui, on utilise abondamment l'échantillonnage PPT des UPÉ dans la conception d'enquêtes à grande échelle, mais on sélectionne dans chaque strate deux ou plusieurs UPÉ sans remise, de sorte que les probabilités d'inclusion des UPÉ sont proportionnelles à la taille (voir la section 5).

Bon nombre d'enquêtes à grande échelle sont répétées au fil du temps, comme l'Enquête sur la population active (EPA) du Canada, menée chaque mois, et la Current Population Survey (CPS) des États-Unis, avec remise partielle des unités finales (appelée aussi échantillonnage par renouvellement). Dans le cas de l'EPA, par exemple, l'échantillon de ménages est divisé en six groupes de renouvellement (échantillons constants ou panels) et un groupe de renouvellement reste dans l'échantillon pendant six mois consécutifs, puis est retiré de l'échantillon, ce qui donne un chevauchement de cinq sixièmes entre deux mois consécutifs. Dans la foulée des travaux initiaux de Jessen (1942) sur l'échantillonnage à deux reprises avec remise partielle des unités, Yates (1949) et Patterson (1950) ont jeté les bases des théoriques de la conception et de l'estimation d'enquêtes à passages répétés et démontrent qu'on pouvait améliorer l'efficacité de l'estimation de niveau et de changement en tirant parti des données antérieures. Hansen, Hurwitz, Nisselson et Steinberg (1955) ont mis au point des estimations plus simples, appelées estimations composites K , applicables aux plans d'échantillonnage stratifié à plusieurs degrés avec échantillonnage PPT au premier degré. Rao et Graham (1964) ont étudié des techniques de remise optimale pour les estimateurs composites K . On a également proposé divers ajouts. On a utilisé des estimateurs composites dans le cas de la CPS et d'autres enquêtes permanentes à grande échelle. Encore récemment, l'EPA du Canada a adopté l'estimation composite, appelée estimation composite par

importants, motivés surtout par des critères d'ordre pratique et d'efficacité. L'article marquant de Cochran (1939) présente plusieurs résultats importants : le recours à l'analyse de variance pour estimer l'amélioration de l'efficacité due à la stratification, l'estimation des composantes de la variance dans l'échantillonnage à deux degrés en vue d'études futures sur un sujet semblable, le choix de l'unité d'échantillonnage, l'estimation par régression sous échantillonnage à deux phases et l'effet des erreurs dans la taille des strates. Dans cet article, Neyman a également proposé le concept de superpopulation : « La population finie doit être considérée comme un échantillon aléatoire d'une population infinie. » Il est intéressant de noter qu'à l'époque, Cochran n'était pas d'accord avec le concept traditionnel de population fixe : « En outre, il est loin d'être réaliste de considérer la population comme un lot fixe de nombres connus. » Cochran a également proposé l'estimation par quotient pour les enquêtes par sondage, mais Laplace (1820) avait déjà utilisé l'estimateur par quotient. Dans un autre article marquant, Cochran (1942) a formulé la théorie de l'estimation par régression. Il a calculé la variance conditionnelle de l'estimateur par régression habituel pour un échantillon fixe ainsi qu'un estimateur échantillon de cette variance, en supposant un modèle de régression linéaire $y = \alpha + \beta x + e$, où e a une moyenne nulle et une variance constante dans les séries statistiques dans lesquelles x est fixe. Il a également noté que l'estimateur par régression restait sans biais (par rapport au modèle) sous échantillonnage non aléatoire, à condition que le modèle de régression linéaire hypothétique soit correct. Il a calculé le biais moyen en cas de d écarts par rapport au modèle (notamment dans le cas de la régression quadratique) pour l'échantillonnage aléatoire simple à mesure que la taille de l'échantillon n augmente. Cochran a ensuite étendu ses résultats à la régression pondérée et calculé le résultat d'optimalité, selon lui, il s'agit de « la meilleure estimation linéaire sans biais si la valeur moyenne et la variance changent proportionnellement à x ». Dans les travaux récents, ce dernier modèle est appelé modèle par quotient. Madow et Madow (1944) et Cochran (1946) ont comparé la variance prévue sous un modèle de superpopulation pour étudier analytiquement l'efficacité relative de l'échantillonnage systématique et de l'échantillonnage aléatoire stratifié. Cet article a incité d'autres chercheurs à mener des travaux sur l'utilisation de modèles de superpopulation dans le choix de stratégies d'échantillonnage probabiliste, ainsi que sur l'inférence dépendante d'un modèle et l'inférence assistée par un modèle (voir la section 3).

En Inde, Mahalanobis a fait un apport innovateur à la théorie de l'échantillonnage en formulant des fonctions de coût et de variance pour la conception d'enquêtes. Son

probabiliste (ou fondé sur le plan de sondage) en ce qui concerne l'inférence à partir d'échantillons d'enquête. Il a montré, avec des arguments théoriques et des exemples pratiques, que l'échantillonnage aléatoire stratifié était préférable à l'échantillonnage équilibré, car ce dernier peut donner de mauvais résultats si les hypothèses sous-jacentes du modèle sont violées. Neyman a également avancé, dans sa théorie de l'échantillonnage aléatoire stratifié sans remise, les notions d'efficacité et de répartition optimale en assouplissant la condition des probabilités d'inclusion égales. En généralisant le théorème de Markov sur l'estimation par les moindres carrés, Neyman a prouvé que la moyenne stratifiée, $\bar{y}_s = \sum_h W_h \bar{y}_h$, était le meilleur estimateur de la moyenne de population, $\bar{Y} = \sum_h W_h \bar{Y}_h$, dans la classe linéaire d'estimateurs sans biais de forme $\bar{y}_b = \sum_h W_h \sum_i b_{hi} y_{hi}$, où W_h, \bar{y}_h et \bar{Y}_h sont le poids, la valeur de l'élément y_{hi} observée au moment du i^{e} tirage d'échantillon ($i = 1, \dots, m_h$) dans la h^{e} strate. On a obtenu la répartition optimale (n_1, \dots, n_L) de la taille de l'échantillon total, n , en minimisant la variance de \bar{y}_s sous réserve de $\sum_h n_h = n$: on a découvert plus tard une preuve antérieure de la répartition de Neyman par Tschuprow (1923). Neyman a également proposé une inférence à partir de grands échantillons en fonction d'intervalles de confiance selon la théorie normale, de manière que la fréquence des erreurs dans les énoncés de confiance en fonction de tous les échantillons aléatoires stratifiés qu'il est possible de tirer n'excède pas la limite prescrite à l'avance « quelles que soient les propriétés inconnues de la population ». Une méthode d'échantillonnage qui satisfait l'énoncé de fréquence susmentionné est dite « représentative ». Il est à noter que Hubback (1927) avait déjà fait allusion à l'énoncé de fréquence associé à l'intervalle de confiance. Dans son dernier apport à la théorie des enquêtes par sondage, Neyman (1938) a étudié l'échantillonnage à deux phases de stratification et calculé la taille optimale des échantillons de première phase et de deuxième phase, n' et n , en minimisant la variance de l'estimateur sous réserve d'un coût donné $C = n'c' + nc$, où le coût par unité de deuxième phase, c , est élevé par rapport au coût par unité de la première phase, c' .

Au cours des années 1930, la demande d'information a connu une croissance rapide et l'on a pris conscience des avantages de l'échantillonnage probabiliste—portée accrue, réduction de coût, plus grande vitesse et caractéristiques indépendantes d'un modèle—, d'où une augmentation du nombre et du type d'enquêtes menées par échantillonnage probabiliste et couvrant de grandes populations. La presque totalité des statisticiens d'enquête ont adopté l'approche de Neyman. En outre, cette dernière a inspiré divers ajouts

Evaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage

J.N.K. Rao¹

Résumé

Une grande partie de la théorie des enquêtes par sondage a été motivée directement par des problèmes d'ordre pratique survenus au moment de la conception et de l'analyse des enquêtes. En revanche, la théorie des enquêtes par sondage a influencé la pratique, ce qui a souvent donné lieu à des améliorations importantes. Dans le présent article, nous examinons cette interaction au cours des 60 dernières années. Nous présentons également des exemples où une nouvelle théorie est nécessaire ou encore où la théorie existe sans être utilisée.

Mots clés : Analyse des données d'enquête; apports antérieurs; question d'inférence; méthodes de rééchantillonnage; estimation sur petits domaines.

1. Introduction

Dans cet article, je vais examiner l'inter-relation entre la théorie des sondages et la pratique dans les quelques 60 dernières années. Je vais couvrir une grande variété de sujets: les premières contributions significatives qui ont grandement influencé la pratique, les questions d'inférence, l'estimation par calage qui assure la cohérence aux totaux établis de variables auxiliaires, l'échantillonnage à probabilités inégales sans remplacement, l'analyse de données d'enquêtes, le rôle des méthodes de ré-échantillonnage, et l'estimation pour petits domaines. Je vais aussi présenter quelques exemples où il y a soit besoin d'une nouvelle théorie soit une théorie existante qui n'est pas tellement utilisée.

2. Quelques apports marquants : 1920 – 1970

La présente section rend compte de certains apports marquants à la théorie et aux méthodes des enquêtes par sondage, apports qui ont grandement influencé la pratique. Le statisticien norvégien A.N. Kjaer (1897) fut sans doute le premier à promouvoir l'échantillonnage (appelé « méthode représentative » à l'époque) plutôt qu'un dénombrement complet, quoique la plus ancienne référence à l'échantillonnage remonte au grand récit épique indien Mahabharata (Hacking 1975, page 7). Dans la méthode représentative, l'échantillon doit refléter la population mère finie; à cette fin, on procède par échantillonnage équilibré au moyen de la sélection raisonnée ou par échantillonnage aléatoire. On utilise la méthode représentative en Russie dès 1900 (Zarkovic 1956) et, vers la même époque, Wright l'employa pour mener des enquêtes par sondage aux États-Unis. Dans les

années 1920, on utilisait abondamment la méthode représentative, et l'Institut international de statistique joua un rôle de premier plan en créant en 1924 un comité chargé de produire un rapport sur cette méthode. Le rapport de ce comité portait sur des aspects théoriques et pratiques de la méthode d'échantillonnage aléatoire. Bowley (1926) contribua à ce rapport par ses travaux fondamentaux sur l'échantillonnage aléatoire stratifié avec répartition proportionnelle, qui permit de tirer un échantillon représentatif avec probabilités d'inclusion égales. Hubback (1927) prit conscience de la nécessité d'un échantillonnage aléatoire dans les enquêtes sur les cultures : « La seule façon d'arriver à une estimation satisfaisante consiste à établir une approximation de l'échantillonnage aléatoire aussi proche que les circonstances le permettent, car ainsi, non seulement on élimine les limites personnelles de l'expérimentateur, mais il devient possible de déterminer la probabilité avec laquelle les résultats d'un nombre donné d'échantillons se situent à l'intérieur d'une étendue donnée par rapport à la moyenne arithmétique. Concrètement, il s'agit de trouver combien d'échantillons sont nécessaires pour assurer que la probabilité soit d'au moins 20:1 par rapport à la moyenne des échantillons à l'intérieur d'un maund de la vraie moyenne. » Cet énoncé contient deux observations importantes concernant l'échantillonnage aléatoire : 1) il évite les biais personnels dans la sélection d'un échantillon; 2) on peut déterminer la taille de l'échantillon pour satisfaire une marge d'erreur spécifiée par rapport à une chance de 1 sur 20. Mahalanobis (1946b) a observé que les travaux fondamentaux de R.A. Fisher sur la conception des expériences, menés à la Rothamsted Experimental Station, furent directement influencés par Hubback (1927). Dans un article marquant, devenu un classique, Neyman (1934) a jeté les bases théoriques de l'échantillonnage

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, (Ontario), Canada, K1S 5B6.

Membres du comité de sélection de l'article Waskberg (2005-2006)

Gordon Brackstone, (Président)
 Wayne Fuller, *Iowa State University*
 Sharon Lohr, *Arizona State University*

Présidents précédents :

Graham Kalton (1999 - 2001)
 Chris Skinner (2001 - 2002)
 David A. Binder (2002 - 2003)
 J. Michael Brick (2003 - 2004)
 David R. Bellhouse (2004 - 2005)

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'engagement de théorie et de pratique caractéristique des travaux de Waksberg. L'auteur reçoit une prime en argent qui provient d'une bourse de Westat, en reconnaissance des contributions de Joe Waksberg pendant ses nombreuses années de collaboration avec Westat. L'administration financière de la bourse est assurée par l'American Statistical Association. Gad Nathan, Wayne Fuller, Tim Holt, Norman Bradburn, Jon Rao et Alastair Scott sont les gagnants précédents. Les cinq premiers articles de la série sont déjà parus dans la revue *Techniques d'enquête*.

Précédents gagnants du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)

Nominations:

L'auteur de l'article Waksberg de 2007 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées à Gordon Brackstone, président du comité, à 78, chemin Charing, Ottawa (Ontario), Canada, K2G 4C9, par courriel à Gordon.brackstone@sympatico.ca ou par télécopieur au (613) 951-1394. Les candidatures et les suggestions de sujets doivent être reçues d'ici au 28 février 2006.

Article sollicité Waksberg 2005

Auteur: J.N.K. Rao

J.N.K. Rao est professeur distingué de recherche à l'Université Carleton d'Ottawa. Il a publié plusieurs articles sur une vaste étendue de sujets en théorie et méthodes de sondages et est auteur du livre de 2003 chez Wiley "Small Area Estimation". Son intérêt pour la recherche en échantillonnage inclut l'analyse de données d'enquêtes, l'estimation pour petites régions, les données manquantes et l'imputation, les méthodes de ré-échantillonnage et l'inférence à l'aide de la vraisemblance empirique. Son article du JASA de 1981 (avec A.J. Scott) sur l'analyse de données d'enquêtes a été sélectionné parmi les articles phares (landmark paper) de la théorie et des méthodes d'échantillonnage. Il est membre du Comité consultatif des méthodes statistiques de Statistique Canada depuis sa création il y a 20 ans. Il est fellow de la Société Royale du Canada et a reçu la médaille d'or de la Société statistique du Canada en 1994.

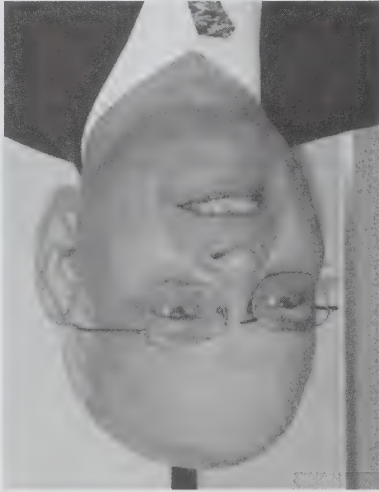
À la mémoire M.P. Singh (1941-2005)

d'un fonds permanent pour la recherche en méthodologie, et il a présidé personnellement le Comité de recherche et de développement en méthodologie à ses débuts. Il a encouragé de nombreux chercheurs et s'est donné beaucoup de mal pour qu'ils se sentent chez eux à Statistique Canada. La soixantaine n'a pas du tout endigué le flot de ses idées. M.P. a déployé une énergie considérable au cours des quatre dernières années pour proposer une révision majeure de la façon d'effectuer les enquêtes sur les ménages au Canada. Comme résultat de ses efforts, on travaille à travers Statistique Canada à des façons de mettre en oeuvre sa vision, et son influence sur les enquêtes des ménages se fera sentir pour des années à venir.

M.P. était spécialement attaché à la recherche en statistique et à la profession statistique. Il était l'auteur de plus de 40 articles dans des revues internationales, le coéditeur de deux livres publiés chez Wiley&Sons, et a organisé plusieurs séances et fait des présentations lors de plusieurs conférences statistiques. Il a siégé sur plusieurs comités et ateliers de travail de la Société statistique du Canada, et de l'American Statistical Association. Il a aussi été secrétaire du comité consultatif de Statistique Canada sur les méthodes statistiques.

En retour, il a reçu les honneurs de la profession: il a été élu à l'International Statistical Institute en 1975, et en 1988 il est devenu Fellow de l'American Statistical Association. Néanmoins, c'est son influence sur toute une génération de statisticiens qui constitue son plus grand héritage. Il a été un mentor, un moniteur, un père et un ami pour tous ceux qui l'ont connu. Il a inspiré les autres à offrir le meilleur d'eux-mêmes, ce qu'ils ont fait. Il était toujours prêt pour un titre, un sourire et un mot amical d'encouragement. Il a consacré sa vie à la profession statistique et c'est à travers ceux qu'il a atteints qu'on peut mesurer sa véritable contribution.

Il laisse dans le deuil son épouse Savitri, ses deux filles Mala et Mamta, et son fils Rahul.



Dr. Mangala P. Singh né en Inde le 26 décembre 1941, il avait obtenu un doctorat de l'Indian Statistical Institute en 1969, avec une spécialisation en échantillonnage d'enquêtes. Il s'était joint à Statistique Canada en 1970, où il avait atteint le poste de directeur de la division des méthodes d'enquêtes des ménages en 1994, poste qu'il a occupé jusqu'à son décès le 24 août 2005.

M.P., comme tout le monde l'appelaient, était une figure de proue pour l'application des méthodes statistiques à Statistique Canada. Il était probablement plus étroitement associé à l'enquête sur la population active, l'une des

enquêtes les plus importantes de l'agence. Il a dirigé la méthodologie de l'enquête sur la population active au cours de plusieurs remaniements durant les années 1970, 1980, 1990 et au début du 21^e siècle, en introduisant à chaque fois des innovations tout en s'assurant que les changements étaient valides et bien testés. Au cours des dernières années de sa carrière, il a veillé au développement de plusieurs enquêtes nouvelles et innovatrices dans le domaine de la santé et au développement de programmes statistiques dans les domaines des dépenses des ménages, de l'éducation et de la justice.

Le rôle de M.P. comme rédacteur en chef de la revue *Techniques d'enquête* a transformé le rôle des techniques d'enquête, à la fois au Canada et à l'étranger. M.P. a été le rédacteur fondateur de la revue, et au cours de 30 années, l'a fait évoluer jusqu'à ce qu'elle devienne une publication amicale de statisticiens. Grâce à sa capacité à attirer un réseau de rédacteurs adjoints et de contributeurs, *Techniques d'enquête* est maintenant reconnue comme une des revues prééminentes dans son domaine à travers le monde. Même au cours des dernières années, M.P. a continué d'innover, comme avec l'introduction de la série d'articles *Waksberg* et de la publication électronique.

M.P. a été une source de plusieurs autres "grandes idées" tout au long de sa carrière à Statistique Canada. Durant les années 1970, il a contribué à gagner des appuis pour l'idée

Belsby, Bjørnstad et Zhang discutent de la modélisation en vue d'estimer le nombre de ménages de diverses tailles en présence de non-réponse non ignorable. Ils modélisent le mécanisme de réponse sachant la taille du ménage, en utilisant la taille enregistrée de la famille comme donnée supplémentaire. Ils décrivent d'abord l'élaboration de leurs méthodes de modélisation, puis produisent et évaluent des estimations à l'aide de données provenant de l'Enquête sur les dépenses de consommation en Norvège de 1992.

Nandram, Cox et Choi considèrent l'analyse de données catégoriques provenant d'un seul tableau à double entrée en présence de non-réponse partielle ainsi que totale ou, selon leur terminologie, de scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ignorabilité. Ils illustrent leurs méthodes à l'aide de données bivariées incomplètement observées provenant de la National Health and Nutrition Examination Survey, où les variables pour lesquelles des données manquent sont la densité minérale osseuse et le revenu familial.

Dans la première de trois notes brèves publiées dans le présent numéro, Beaumont discute de l'utilisation de l'information sur le processus de collecte des données lors de la correction de la pondération pour tenir compte de la non-réponse. Puis, il donne un exemple tiré de l'Enquête sur la population active du Canada en utilisant le nombre de tentatives de prise de contact avec une unité étudiée. Un résultat important est que, si l'information sur le processus de collecte peut être considérée comme étant aléatoire, la méthode n'introduit aucun biais.

En partant de principes fondamentaux, Bustos dérive une forme explicite de la fonction de probabilité d'un échantillon ordonné. Puis, il montre comment on peut l'utiliser pour calculer les probabilités d'inclusion et offre des exemples pour des plans de sondage courants. Enfin, il donne la forme générale de la matrice des corrélations des unités d'échantillonnage, qui dépend uniquement des probabilités d'inclusion.

Enfin, dans son article, Wu passe brièvement en revue certains aspects théoriques de la méthode de pseudo-vraisemblance empirique en échantillonnage et présente des algorithmes permettant de calculer l'estimateur du maximum de pseudo-vraisemblance empirique et de construire les intervalles de confiance des rapports de pseudo-vraisemblance empirique. Il donne des fonctions utilisant les logiciels statistiques R et S-PLUS pour faciliter l'implémentation de ces algorithmes dans le cas d'enquêtes réelles ou d'études en simulation.

Harold Mantel

Dans ce numéro

C'est avec une profonde tristesse que nous annonçons le décès récent de M.P. Singh, rédacteur en chef de la revue *Techniques d'enquête* depuis la publication du tout premier numéro en 1975. Le présent numéro débute par un bref article nécrologique en sa mémoire.

Ce numéro de *Techniques d'enquête* contient aussi le cinquième article de la série d'articles annuels sollicités à Joseph Waksberg. Une brève biographie de ce dernier est parue dans le numéro de juin 2001 de la revue, en même temps que le premier article de la série. Je tiens à remercier les membres du comité de sélection – Michael Brick, président, David Bellhouse, Gordon Brackstone et Paul Biemer – d'avoir choisi Jon Rao comme auteur de l'article Waksberg de cette année.

Dans son article intitulé « Interaction entre la théorie et la méthodologie des enquêtes par sondage : Une évaluation », Rao montre comment les progrès théoriques stimulent l'élaboration de méthodes d'enquête et comment la pratique des enquêtes force à remettre la théorie en question. Il commence par résumer 50 années de contributions, de 1920 à 1970, puis présente une discussion plus approfondie de faits nouveaux récents dans plusieurs domaines. Enfin, il donne plusieurs exemples de théories importantes qui ne sont pas encore appliquées à grande échelle en pratique.

Dans leur article, Fuller et Kim élaborent et étudient une méthode d'imputation hot-deck efficace sous l'hypothèse que les probabilités de réponse sont égales dans les cellules d'imputation. La méthode qu'ils proposent est fondée sur la notion d'imputation fractionnaire et s'appuie sur des techniques de régression pour obtenir une approximation de la version entièrement efficace de l'imputation fractionnaire. Ils élaborent une estimation de la variance pour les méthodes de rééchantillonnage et montrent que la méthode qu'ils proposent donne de bons résultats dans une étude en simulation.

L'article de Brick, Jones, Kalton et Valliant décrit la comparaison, au moyen d'une étude en simulation, de trois méthodes d'estimation de la variance en présence d'imputation hot-deck, à savoir la méthode assistée par modèle, la méthode du jackknife corrigée et la méthode d'imputation multiple. Le but de l'étude en simulation est d'étudier les propriétés de ces estimateurs de la variance quand les hypothèses sous-jacentes ne sont pas vérifiées. Les auteurs constatent que le taux de couverture des intervalles de confiance ne s'approche pas du niveau nominal quand les estimations ponctuelles sont biaisées parce que l'on omet de tenir compte des domaines d'intérêt à l'étape de l'imputation. Ils concluent en notant que les différences entre les estimateurs de la variance sont trop faibles et incohérentes pour qu'on puisse affirmer que l'un d'entre eux est supérieur aux autres en général.

Littlé et Varthavian étudient l'effet de la pondération pour la non-réponse sur l'erreur quadratique moyenne (EQM) d'un estimateur de la moyenne de population. Ils corrigent la pondération pour tenir compte de la non-réponse en ajustant les poids de sondage au moyen de l'inverse des taux de réponse dans les cellules. Ils concluent que, pour réduire le biais de non-réponse, une covariable de pondération doit avoir deux caractéristiques : elle doit être corrélée à la probabilité de réponse, d'une part, et à la variable de résultat, d'autre part. Si cette deuxième caractéristique est vérifiée, la pondération peut aussi réduire la variance due à la non-réponse. Ils proposent des estimations de l'EQM et les utilisent pour définir un estimateur composite. Celui-ci donne de bons résultats lors d'une évaluation par étude en simulation.

O'Malley et Zaslavsky présentent des modèles de fonctions de variance et de covariance généralisées (FVCG) pour des moyennes multivariées de questions d'enquête ordonnées, dans le cas de données complètes ainsi que de données avec non-réponse structurée. Ils commencent par décrire l'élaboration et l'évaluation de leurs méthodes, puis ils illustrent ces dernières à l'aide de données provenant de la Consumer Assessment of Health Plans Study. Dans la conclusion, ils discutent de certaines questions liées à l'application des FVCG.

L'article de Singh, Shukla et Kundu décrit l'élaboration de modèles spatiaux et spatio-temporels pour l'estimation sur petits domaines, ainsi que pour l'estimation de l'erreur quadratique moyenne du meilleur prédicteur linéaire sans biais empirique (EBLUP) résultant. Ils appliquent leurs modèles aux données sur les dépenses de consommation mensuelles par habitant et concluent qu'ils peuvent être très efficaces s'il existe des corrélations importantes dues aux effets de quartier.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 31, numéro 2, décembre 2005

Table des matières

Dans ce numéro	121
À la mémoire de M.P. Singh	123
Article Sollicite Waksberg	
J.N.K. Rao	127
Articles Réguliers	
Wayne A. Fuller et Jae Kwang Kim	153
Imputation hot deck pour le modèle de réponse	153
J. Michael Brick, Michael E. Jones, Graham Kalton et Richard Valliant	165
Estimation de la variance avec imputation hot deck : Une étude par simulation de trois méthodes	165
Roderick J. Little et Sonya Varivartan	175
La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage?	175
Alistair James O'Malley et Alan Mark Zaslavsky	185
Fonctions de variance-covariance pour les moyennes de domaine des questions avec valeurs ordonnées	185
Bharat Bhusan Singh, Girya Kant Shukla et Debasis Kundu	201
Modèles spatio-temporels pour l'estimation pour petits domaines	201
Liv Belsby, Jan Bjørnstad et Li-Chun Zhang	215
Méthodes de modélisation et d'estimation de la taille du ménage en présence de non-réponse non ignorable appliquées à l'Enquête sur les dépenses de consommation de la Norvège	215
Balgebin Nandram, Lawrence H. Cox et Jai Won Choi	233
Analyse bayésienne des données catégoriques manquantes non ignorables : Une application à la densité minérale osseuse et au revenu familial	233
Communications brèves	
Jean-François Beaumont	249
L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids	249
Alfredo Bustos	255
Structure de corrélation des unités d'échantillonnage	255
Changbao Wu	261
Algorithmes et codes R pour la méthode de la pseudo-vraisemblance empirique dans les sondages	261
Remerciements	267

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président D. Royce
Anciens présidents G.J. Brackstone
R. Platak
E. Rancourt (Gestionnaire de la production)
D. Roy
M.P. Singh

COMITÉ DE RÉDACTION

Rédacteur en chef
Rédacteur en chef délégué M.P. Singh, *Statistique Canada*
H. Mantel, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistique Canada*
J.M. Brick, *Westat, Inc.*
P. Cantwell, U.S. Bureau of the Census
J.L. Eltinge, U.S. Bureau of Labor Statistics
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Office for National Statistics*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
J. Kovar, *Statistique Canada*
P. Lahiri, *JPSM, University of Maryland*
G. Nathan, *Hebrew University*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J.-F. Beaumont, P. Dick et W. Yung, *Statistique Canada*

J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Schuren, *National Opinion Research Center*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
R. Tillie, *Université de Newchâtel*
Y. Valliant, *JPSM, University of Michigan*
V.J. Verma, *Università degli Studi di Siena*
J. Wasberg, *Westat, Inc.*
K.M. Wolter, *Iowa State University*
A. Zaslavsky, *Harvard University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (smj@statcan.ca, Division des instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 30 \$ CA (15 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Février 2006

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrément sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2006

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2005 • VOLUME 31 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 2

•

VOLUME 31

•

DÉCEMBRE 2005

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE

